DYNAMIC SPEECH IMAGING WITH LOW-RANK APPROXIMATION

BY

MAOJING FU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Adviser:

Professor Zhi-Pei Liang

ABSTRACT

Dynamic speech imaging is a powerful technique for real-time visualization of speech dynamics. As a promising modality for dynamic speech imaging, magnetic resonance imaging (MRI) can provide good soft-tissue contrast in an arbitrary imaging plane with a non-invasive procedure. However, conventional MRI suffers from low spatiotemporal resolution, which limits its application in dynamic speech imaging. This thesis presents a novel model-based dynamic MR imaging method to capture speech dynamics in high spatiotemporal resolution.

Specifically, high spatiotemporal resolution reconstruction from very sparsely sampled data is achieved using the partial separability (PS) model, which takes advantage of the spatiotemporal correlations of dynamic speech images. The sampling pattern is also optimized to better capture speech dynamics. The spatial-spectral sparsity constraint is further incorporated into the basic PS model-based reconstruction to improve reconstruction quality. The effectiveness of the above approaches is demonstrated through systematic simulations and preliminary *in vivo* experiments.

To Dad, Mom and Meng

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to my adviser, Professor Zhi-Pei Liang, for all the support, guidance and patience that he has given me over the past two years. The inspiration and advice from him are always very enlightening and have fundamentally reshaped my way of thinking, study and research.

I am also grateful to Professor Brad Sutton for his support and guidance since I started my study at the University of Illinois and for providing me the opportunities to explore the exciting realm of dynamic speech imaging and high performance computing.

I would like to thank my excellent group members, Bo Zhao, Chao Ma and Fan Lam for creating a productive and cooperative working atmosphere in our group and for spending a lot of time and effort giving me advice and helping me revise drafts of this thesis. I am also thankful to my other group members, Anthony Christodoulou, Justin Haldar, Hien Nguyen, Xiaobo Qu, Xi Peng and Huiqian Du for their help and useful discussions during thesis research.

I would like to thank my parents for their unfailing love, support and encouragement. Without their support and encouragement, I would not have the confidence and perseverance to go through this journey. I would also like to thank my girlfriend, Meng, and her family for their understanding and encouragement in times of difficulties.

This thesis would not have been possible without my friends. I would also like to thank all my friends who have given me help and support to complete this thesis. I am deeply indebted to David Ho, Tan Nguyen, Joe Holtrop, Kerui Min, Aiguo Han, Shiyu Chang, Aolin Xu, Alex Pawlicki, Sujeeth Bharadwaj, Chong Li, Roger Serwy, Adam Luchies and Panying Rong for their valuable assistance and suggestions.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

CS	Compressed sensing / Compressive sensing
DFT	Discrete Fourier transform
FFT	Fast Fourier transform
FLASH	Fast low-angle shot
FMRI	Functional MRI
FOV	Field of view
GRAPPA	Generalized autocalibrating partially parallel acquisition
k-t BLAST	k-t broad-use linear acquisition speedup technique
k-t FOCUSS	k-t space focal underdetermined system settler
k-t SENSE	k-t sensitivity encoding
MR	Magnetic resonance
MRI	Magnetic resonance imaging
PARADIGM	Patient adaptive reconstruction and acquisition dynamic imaging method
PARADISE	Patient adaptive reconstruction and acquisition dynamic imaging with sensitivity encoding
PS	Partially separability model
ROI	Region of interest
SENSE	Sensitivity encoding
SMASH	Simultaneous acquisition of spatial harmonics
SNR	Signal-to-noise ratio
SVD	Singular value decomposition
T_E	Echo time
T_R	Repetition time
UNFOLD	Un-aliasing by Fourier-encoding the overlaps using the temporal dimension

CHAPTER 1

INTRODUCTION

1.1 Motivations

Dynamic speech imaging is a powerful tool for real-time visualization of the vocal articulators and their affiliating vocal muscles. Dynamic speech imaging has been applied to analyze a variety of articulator defects and speech disorders, such as cleft palate [1], nasal emission [2], hypernasality [3] and velopharyngeal insufficiencies [4]. However, the state-of-the-art dynamic imaging methods still suffer from low spatiotemporal resolution. Developing advanced dynamic imaging techniques that can significantly improve the spatiotemporal resolution of the dynamic speech images is critical towards deeper understanding of basic speech studies and better clinical diagnostics.

Ideally, an effective dynamic speech imaging should provide: 1) non-invasive procedure to prevent interruption of natural speech; 2) good soft-tissue contrast to clearly visualize the vocal articulators and the affiliating vocal muscles; and 3) high spatiotemporal resolution to capture the fine features and fast motion of the vocal articulators. High spatiotemporal resolution is especially challenging. For instance, visualization of the tongue tip requires a spatial resolution better than 2 mm [5]; and sound production of some fricatives takes less than 75 ms to complete [6], which requires an effective temporal resolution of at least 26 frames per second.

Compared with other imaging methodologies, such as computerized tomography, video fluoroscopy and ultrasound, magnetic resonance imaging (MRI) has the unique capability to capture the speech dynamics with excellent soft-tissue contrast and high spatial resolution and without ionizing radiation risks. While conventional MRI suffers from low temporal resolution, recent development in MRI hardware and image reconstruction techniques has effectively refined the temporal resolution and enabled significant progress in various dynamic imaging applications, such as cardiac MRI and functional MRI. This thesis focuses on applying an optimized data acquisition scheme and an advanced MRI image reconstruction method to improve the spatiotemporal resolution of dynamic speech imaging.

1.2 Problem statement

This thesis attempts to improve the spatiotemporal resolution of dynamic speech imaging with a Partial Separability (PS) model-based imaging method. The Partial Separability (PS) model-based imaging method has been successfully applied to cardiac MRI imaging and achieved up to 2000 times imaging speed acceleration [7]. Therefore, the PS model has great potential to improve the spatiotemporal resolution for other dynamic MRI imaging applications. The hypothesis of this work is that the PS model-based imaging method can effectively improve the spatiotemporal resolution of dynamic speech imaging. Systematic simulations and preliminary *in vivo* experiments are carried out to test this hypothesis.

The PS model assumes that the dynamic object $\rho(\boldsymbol{x}, t)$ is partially separable to the L^{th}-order [8],

$$\rho(\boldsymbol{x}, t) = \sum_{l=1}^{L} U_l(\boldsymbol{x}) V_l(t), \tag{1.1}$$

where the spatial basis function $U_l(\boldsymbol{x})$ describes spatial variations of the imaging object, and the temporal basis function $V_l(t)$ describes temporal variations of the imaging object. Correspondingly, the data acquisition scheme consists of two parts, an imaging data set that captures spatial variation, and a navigator data set that captures temporal variation. Although acquisition

schemes of the imaging data set have been previously investigated [9], the navigator data acquisition schemes have not been fully explored. This thesis attempts to find an optimized navigator sampling pattern to capture more accurate speech dynamics.

Recent development of the PS model-based imaging method includes incorporating sparsity constraint in the image reconstruction and has shown benefits in cardiac MRI applications [10]. This thesis aims to investigate whether this sparsity constraint can lead to better reconstruction for dynamic speech imaging.

1.3 Summary of contributions

The contributions of this thesis are threefold. Firstly, the Partial Separability (PS) model has been employed to improve the spatiotemporal resolution for dynamic speech imaging. Simulations and experimental results have demonstrated that the PS model can capture articulator motion with high spatiotemporal resolution.

Secondly, this thesis has explored multiple navigator sampling patterns in terms of their abilities to capture accurate articulator motion. Specifically, a numerical speech phantom has been developed for the simulations of navigator sampling patterns. Reconstructions based on these sampling patterns were systematically compared in terms of qualitative and quantitative metrics. A spiral navigator has been shown to capture more accurate speech dynamics.

Thirdly, this thesis has applied an existing PS model-based imaging method to improve reconstructions in dynamic speech imaging. This imaging method improves fine features and temporal dynamics of the basic PS reconstruction by imposing partial separability and spatial-spectral sparsity. Improvements in reconstruction quality have been demonstrated with both numerical simulations and *in vivo* experiments.

1.4 Organization of the thesis

This thesis is organized as follows: Chapter 2 gives a brief introduction to non-imaging-based and imaging-based methods for speech studies with an emphasis on reviewing the development of MR-based dynamic speech imaging. Chapter 3 describes the proposed ***(k**, t)*-space imaging method for dynamic speech imaging. It starts with a brief review of ***(k**, t)*-space imaging and then a detailed discussion of Partial Separability (PS) model-based imaging method. This chapter focuses on the Partial Separability (PS) model-based methods. Chapter 4 shows both simulation and experimental results illustrating the performance of the proposed method. Conclusions of this thesis are given in Chapter 5.

CHAPTER 2

BACKGROUND

2.1 Traditional techniques for speech investigation

Speech is a vital body function that affects the quality of life. It serves as a basic means of verbal communication through which a speaker's message can be perceived and understood by a listener [5]. Despite its fundamental importance, speech can also have complex variations and has been a continuing research topic for multiple research areas, including linguistics, acoustics, pathology, speech imaging and speech signal processing [5].

Research in speech-related research areas can be generally divided into two broad categories: speech production and speech perception [5]. Speech production mainly studies the anatomical structure of vocal tract articulators, such as the lips, the tongue, the velum, as well as how these vocal tract articulators mutually interact to produce meaningful sound [5]. Figure 2.1 depicts a midsagittal view of the region of interest in speech production research. Major speech articulators in the upper-airway vocal tract are indicated with white arrows when the speech subject holds a closed-mouth position. Unlike speech production, speech perception investigates how the human brain perceives and translates auditory signal [5]. Research in speech perception attempts to infer the perception process in the human brain by analyzing subtle difference in articulator motion. While research works in speech production and speech perception vary in their specific research attempts, their conclusions are usually based on observation of articulator motion in the

oropharyngeal region. It is important for both research areas, therefore, to be equipped with effective techniques to quantitatively measure subtle difference in speech dynamics. In general, two categories of approaches exist for the study of articulator motion, the imaging-based and non-imaging-based methods.

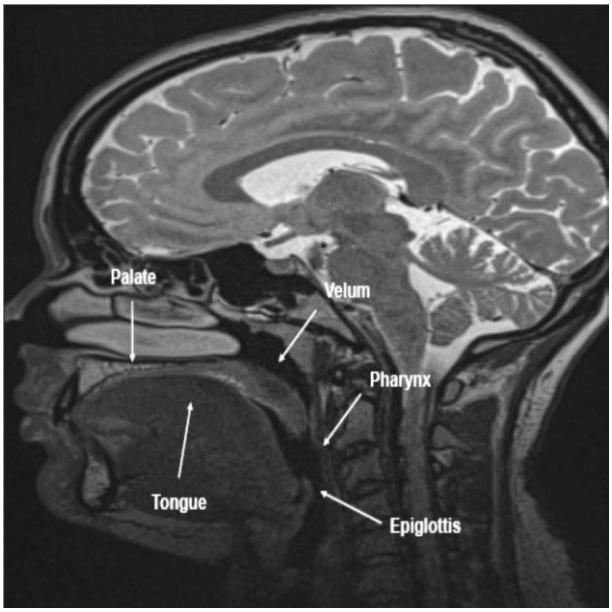


Figure 2.1: Mid-sagittal view of the region of interest in speech research.
(Image courtesy of Dr. Brad Sutton and Dr. Jamie Perry)

2.1.1 Non-imaging-based approaches

The non-imaging-based approach, as its name suggests, applies non-imaging modalities to collect speech related signal. By analyzing the collected signal, one can indirectly infer the physiological state of the speech production system. Examples of the non-imaging approach include electropalatography (EPG), electroglottography (EGG), electromyography (EMG), electromagnetic articulography (EMA) and high-resolution manometry (HRM). EMG, EMA and HRM are three representative methods among these approaches. EMG measures the electric potential, amplitude and timing of the physiological signal in the muscle cells [11]. These parameters can provide information about the strength and duration of muscle contraction. Unusual patterns in mus-

cle contraction may serve as signs of neuropathies, myopathies and neuromuscular junction diseases [11]. EMA evaluates the motion of speech articulators by attaching tiny receiver coils on their surfaces [12]. A magnetic field is placed around the attached receiver coils to allow generation of electric signal when these receiver coils move. The electric signal can then be processed and interpreted as trajectories of articulator movement. Previous research in EMA has already enabled dynamic assessment of tongue function in patients with dysarthria, i.e., impaired ability to produce normal articulation [12]. Assessment of dysarthria is valuable for the understanding of defects in speech motor function and EMA serves as an important tool to evaluate motor functions [12]. HRM usually examines speech muscle contraction with an array of flexible catheters connected to pressure transducers. These transducers can measure amplitude and timing of the contractile wave within a catheter array [13]. Recent developments in HRM have enabled diagnosis of lower esophageal sphincter (LES) relaxation errors [13]. Although there exists other effective non-imaging-based methods for speech studies and it is not possible to provide a complete list of them, the above-mentioned non-imaging-based methods have already effectively enriched speech researchers' ability to characterize speech function.

2.1.2 Imaging-based approaches

The imaging-based approach, however, enjoys twofold advantages when compared with its non-imaging-based counterpart. Firstly, the imaging-based approach can effectively distinguish local articulatory behaviors from global articulator movement in the vocal tract [14]. The vocal tract is usually regarded as a complex physiological system consisting of hard tissue structures (the jaw, the hard palate and the pharyngeal wall) and soft-tissue structures (the tongue tip, the soft palate, the tongue dorsum, the velum and the epiglottis) [14]. Each individual structure in this complex physiological system displays its own movement pattern and couples with other structures to move in a highly correlated manner. For the non-imaging-based approach, it is usually difficult to differentiate the overall correlated articulator motion with the motion of each

individual articulator. This can be exemplified by the velum-pharynx system. While the velum body and the velum tip both contact the pharyngeal wall during speech production, they behave differently in how each of them approaches the pharyngeal wall. By simply analyzing a few parameters provided by non-imaging-based techniques, it is difficult to infer how the bulk of velum contacts the pharynx. With imaging-based techniques, however, it is straightforward to understand how these structures move individually and collectively. Secondly, the imaging-based approach excels at providing direct visualization of the region of interest. Compared with the abstract parameters provided by non-imaging-based methods, an image is a more straightforward way to indicate structural deformation and articulatory deficits in speech analysis. With accurate visualization of the speech dynamics, the imaging-based approach can better facilitate speech analysis [15 - 18]. This can be exemplified by clinical applications of the imaging-based techniques to the diagnosis of sleep apnea [15], motor dysfunction [16], swallowing difficulty [17], and velopharyngeal insufficiency [18]. Research applications of the imaging-based approach include modeling the mechanism for fricative consonants [19]. Four imaging modalities have been traditionally employed to perform dynamic imaging-based experiments. These imaging modalities include video endoscopy, ultrasound, video fluoroscopy and computerized tomography (CT). Each of these imaging modalities has its unique imaging principles and properties.

2.1.2.1 Video endoscopy

The video endoscopy is by its nature an optical imaging modality that visualizes the interior of the human body cavity with a light source [20]. This is usually done by directly inserting a flexible endoscope into the region of interest in the vocal tract. Specifically, the endoscope is often inserted through the oral cavity or nasal cavity. At the front end of the inserted endoscope, a camera with lighting device usually observes tissue color, articulator structure and micro vessel distribution and before it sends back imaging information through a flexible optical fiber [20]. Figure 2.2 depicts a typical image of the pharynx obtained by video endoscopy.

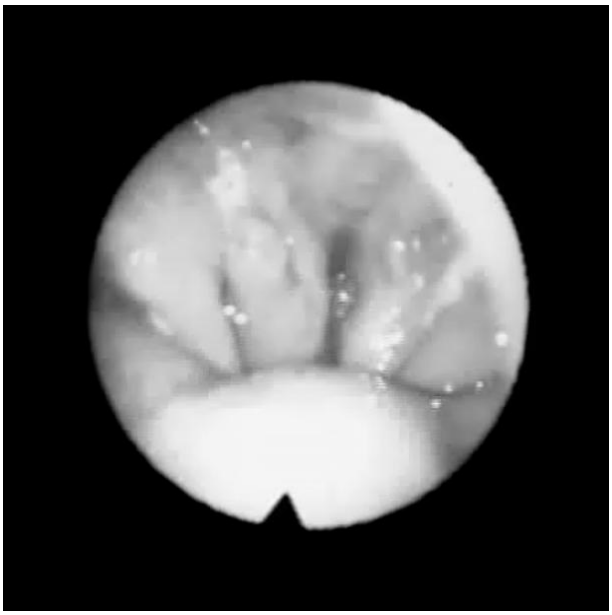


Figure 2.2: A typical video endoscopy image of the pharynx
(Image courtesy of Dr. Brad Sutton and Dr. Jamie Perry)

Although video endoscopy has a simple structure and simple imaging principle, it proves to be a useful tool in the detection of the surface lesions in the vocal tract. Malignant tissues and abnormal micro vessels in the vocal tract mucosa can be easily captured by comparing their shapes and colors with normal patterns [20]. Examples of video endoscopy examination include the assessment of vibratory and mucosal displacement [20] and the evaluation of voice prosthesis complications [21]. Recent development of video endoscopy also results in some variants that serve specific clinical purposes, such as panendoscopy and laryngoscopy [20, 21].

The limitations of video endoscopy mainly come from the use of non-penetrating light. The imaging information obtained with video endoscopy is usually constrained to the surface view within a localized region [20]. Three-dimensional structures, especially soft-tissue articulators with obscuring moving structures below the surface, are difficult to observe and measure with video endoscopy. Also, the application of video endoscopy to speech study is limited by its invasive nature. Patients in an endoscopy examination suffer from a great deal of discomfort since the endoscope tubes are usually inserted into the nasal cavity. Therefore, the application of endoscopy in dynamic speech imaging is not universal.

2.1.2.2 Ultrasound

Unlike video endoscopy, ultrasound provides a less invasive way to explore the oropharyngeal structure. Instead of inserting an imaging instrument into vocal tract cavities, an ultrasound transducer focuses high-frequency or even ultrahigh-frequency sound waves into a sound beam before it injects this sound beam into the human body [22]. The injected sound beam traverses through soft-tissue structures and bounces back when it approaches boundaries between air, tissues and bones [22]. Usually the sound beam returns with a small proportion of its original energy. This energy can be captured by the transducer and then converted into voltage signals [22]. By calculating the timing and amplitude between sound production and reception, the location of soft-tissue articulators can be analyzed and determined. A multichannel transducer is often used when an imaging plane needs to be imaged. The non-invasive nature of ultrasound renders it a promising technique for dynamic speech imaging.

Ultrasound has been widely applied to image vocal tract structure and tongue movement [23]. Since ultrasound has a limited imaging depth due to signal loss, the transducer is often placed between the lower chin and the upper glottis to fully utilize its imaging range. With this method, speech movement in the lower tongue surface and the upper tongue surface can be effectively captured. Previous researchers have applied ultrasound to analyze 2D tongue movement [22], reconstruct 3D tongue surface [23], model 3D tongue movement [24], assist language pathology study [25] and detect gastro-oesophageal carcinoma [26]. For tongue-movement-related applications, ultrasound successfully carries out the experiment with satisfactory imaging speed.

Although ultrasound is a non-invasive imaging modality without significant biohazards, it suffers from two drawbacks [22]. First, ultrasound has decreased imaging range when it encounters a mixture of soft-tissue types. Energy loss of ultrasound due to both reflection and diffraction prevents ultrasound from probing deeper soft-tissue structures, such as the tongue tip and the velum. Secondly, ultrasound has lower soft-tissue contrast when it is compared with other imaging modalities. Therefore, ultrasound is not an ideal imaging method for speech research.

2.1.2.3 Video fluoroscopy and computerized tomography

Compared with ultrasound, video fluoroscopy and computerized tomography (CT) can be applied without worrying about reflections and signal loss at soft-tissue boundaries. Video fluoroscopy is often referred to as "motion X-ray" that combines conventional X-ray fluoroscopy with real-time visualization techniques to provide a series of X-ray images [27]. An X-ray beam travels through the human body from one side and meets an X-ray receiver on the other side with part of its energy absorbed by all types of tissues along the beam path [27]. The received X-ray signal carries information about tissue structures and can be reconstructed as two-dimensional projections of a three-dimensional human body structure [28].

Video fluoroscopy has been widely applied to dynamic speech imaging due to the popularity of existing X-ray imaging systems [27]. Previous research has exploited video fluoroscopy to develop vocal tract models [27], monitor speech disorders [28], explore swallow process [29] and investigate dysphagia in stroke patients [30]. These applications have made video fluoroscopy the "gold standard" imaging modality for many speech experiments. However, the imaging principle of the video fluoroscopy determines that it is not an ideal imaging modality for the articulator motion analysis since it dependents heavily on the analysis of X-ray attenuation, which is significant larger for bones than for soft-tissues [27]. The projected image, therefore, is more sensitive to the overlying bone structures but less sensitive to soft-tissue structures beneath the bones [27]. This can be demonstrated with Fig. 2.3 that depicts a midsagittal view of the vocal tract obtained from video fluoroscopy. As can be seen in the image, jaw bones can severely obscure important movements of soft-tissue articulators beneath them.

Compared to video fluoroscopy, CT has stronger ability to provide soft-tissue contrast and has a dissimilar strategy to encode spatial information of the object [31]. Instead of projecting the imaging slice from a fixed angle, CT obtains projections of the imaging slice from multiple projection angles [31]. Obtaining data from more than one projection angle allows the ensemble of the acquired data to be combined and reconstructed as a two-dimensional image according to the

projection slice theorem [31]. While soft-tissue structures and hard tissue structures are overlapped in X-ray, these structures can be distinctive in CT. With this unique feature, CT is gradually replacing video fluoroscopy in imaging oropharyngeal structures. Recent developments in CT have even replaced the traditional X-ray spiral CT with electron beam CT that yields higher spatiotemporal resolution and imaging speed [31]. These developments provide the speech research community with powerful tools to analyze speech dynamics.

While video fluoroscopy and CT provide high imaging speed and high spatial resolution, their usage in studying speech dynamics is limited by the use of ionizing radiation [31]. With both modalities, it is not feasible to carry out large-scale study of normal speech and swallowing function since the speech subjects are inevitably exposed to excessive ionizing radiation.

Figure 2.3: A typical video fluoroscopy image of the upper vocal tract
(Image courtesy of Dr. Brad Sutton and Dr. Perlman)

2.2 MR-based dynamic speech imaging

Dynamic MRI serves as a powerful technique to image speech dynamics in full three dimensions [10]. As a non-ionizing imaging modality, MRI measures proton density of soft-tissue structures from arbitrary imaging planes across the entire vocal tract. These preferable characteristics of MRI give it special advantage to serve as a dynamic speech imaging modality. Despite its potentials, the application of MRI to capture the spatiotemporal speech dynamics is usually limited by low imaging speed [10]. Therefore, a lot of research has focused on capturing the spatiotemporal features of natural speech production with limited MR imaging speed.

Early applications of dynamic MRI to speech imaging have been largely constrained to prolonged utterance [32]. Although MR imaging speed is slow compared to the natural movement of vocal tract articulators, it is possible to hold a specific articulator motion within a period of time so that the original dynamic imaging problem can be reduced to an easier structural scan problem. Articulator motion observed with this method is usually produced by simple vowels, such as /a/ and /i/. Under this simplified version of normal articulations, researchers were able to identify correlation between articulator shaping and speech characteristics. Examples include the evaluation of functional positions for the tongue and soft palate [32], and the investigation of a vocal tract transfer function between spatial position and speech frequencies [33]. Despite these successes, the bottleneck of this "prolonged utterance" strategy lies in the fact that only one static image is sampled among the entire speech production process and no transition between natural articulations can be captured [33]. With only one single image, subconscious transitions that cannot be arbitrarily controlled by the speech subject can hardly be captured [33]. Therefore, the early success in prolonged utterance had inspired researchers to capture the entire articulations in a more natural manner.

Speech gating techniques have been later proposed to obtain the "averaged articulator motion" by acquiring ***k***-space data across multiple repetitions [34]. In order to merge the gap between fast-changing articulator motion and limited imaging speed, the gating technique was

proposed to enable sampling ***k***-space data from multiple repetitions of periodic utterance [34]. The feasibility of the gating technique is deeply rooted in two fundamental assumptions. Firstly, the MR acquisition speed only allows a small number of ***k***-space lines to be sampled within any time instance. Secondly, articulator motion across all repetitions is assumed to be identical. Specifically, in each individual repetition, non-overlapping ***k***-space lines are chosen. At the end of all repetitions, the entire ***k***-space is fully covered by using data from each repetition. In this manner, acquiring ***k***-space data from multiple repetitions is equivalent to instantaneous sampling the entire ***k***-space in one repetition. Notable applications of this technique include the detection of cleft palate movement at a speed of 30 frames per second [34], as well as the discovery of complex tongue-palate interaction with 30 ms temporal resolution [35].

While the speech gating technique effectively increases the temporal resolution of speech experiments, the resultant spatiotemporal resolution is subject to the speaker's ability to reproduce consistent articulation motion across multiple repetitions [36]. Motion artifacts will appear in the reconstructed image even when slight motion variation exists among some of the repetitions. Previous research has demonstrated that slight motion variability may heavily corrupt spatiotemporal dynamics for even simple utterance, such as "golly" [36]. Moreover, this technique requires accurate synchronization for each repetition [35]. Mismatch between the speech articulation and indexing timing would also severely degrade reconstruction [35]. These characteristics of the gating technique switched researchers' attention to real-time speech acquisition.

2.3 Challenges in visualizing articulatory dynamics

Recent research on dynamic speech MRI has mainly focused on real-time acquisition. The term "real-time" in this thesis doe not refer to simultaneous data acquisition and image reconstruction. Rather, "real-time" merely suggests "non-gated acquisition". Natural speech variations across repetitions of speech samples are allowed in real-time acquisition. Instead of acquiring the "average articulator motion" in speech gating techniques, the articulator motion in real-time acquisi-

tion does not need to be periodic at all.

Real-time acquisition of speech dynamics has high demand on both data acquisition and image reconstruction. The low-speed nature of MR acquisition continues to limit real-time data acquisition of rapidly changing articulator motion [37]. Conventionally, acceleration in imaging speed is realized through two approaches: specialized hardware and advanced pulse sequences. These approaches include the design of fast pulse sequences and efficient sampling trajectories to traverse ***k***-space [37 - 39], as well as the use of parallel imaging techniques to speed up image acquisition [40 - 42]. An example of these approaches is the implementation of a spiral FLASH sequence to sample the ***k***-space at high imaging speed [43, 44]. With this advanced FLASH sequence, the imaging speed is sufficient to observe some periodic stop consonants [45], which usually requires an effective temporal resolution of 10 ~ 20 frames per second. However, recent development of dynamic speech applications requires real-time visualization of speech dynamics in multiple imaging planes [46]. Visualization of speech dynamics simultaneously in the midsagittal, the coronal and the axial planes would require higher acceleration in imaging speed that conventional methods can hardly offer. Therefore, further acceleration needs to be enabled by taking advantage of the spatiotemporal correlation of the desired speech dynamics. Chapter 3 will discuss in detail how further acceleration in imaging speed is achieved.

CHAPTER 3

PROPOSED METHOD

3.1 Sparse sampling in *(k, t)*-space

3.1.1 Limits on *(k, t)*-space Nyquist sampling

In static MR experiments, the relationship between the measured data, $d(\mathbf{k})$, and the targeted image function, $\rho(\mathbf{r})$, can be expressed as a Fourier transform relationship,

$$d(\boldsymbol{k}) = \int \rho(\boldsymbol{r}) e^{-i2\pi \boldsymbol{k}\cdot\boldsymbol{r}} d\boldsymbol{r}, \tag{3.1}$$

where $\mathbf{r} = (x, y)^{\mathrm{T}}$, x and y denote the two-dimensional spatial coordinates, $\mathbf{k} = (k_x, k_y)^{\mathrm{T}}$, k_x and k_y denote the two-dimensional ***k***-space coordinates. In dynamic MR experiments, however, due to the time-varying nature of the objects being imaged, this ***k***-space description alone is inadequate in expressing temporal events. Instead, the *(**k**, t)*-space has been proposed as an extended description to encompass both spatial and temporal variation [47]. Under the *(**k**, t)*-space description, the measured data can be extended as $d(\boldsymbol{k}, t)$ and the imaging equation can be rewritten as,

$$d(\boldsymbol{k}, t) = \int \rho(\boldsymbol{r}, t) e^{-i2\pi \boldsymbol{k}\cdot\boldsymbol{r}} d\boldsymbol{r}, \tag{3.2}$$

where t denotes the imaging time frames. If $\rho(\boldsymbol{r}, t)$ is support limited, it may seem that an ideal image can be obtained as long as the *(**k**, t)*-space is sampled at the spatiotemporal Nyquist rate.

In practice, however, it is usually difficult to attain the Nyquist sampling rate simultaneously in space and time. This is because the temporal sampling speed that satisfies temporal Nyquist rate often leads to low spatial resolution and *vice versa*.

An intrinsic compromise exists between the spatial resolution and the temporal resolution in dynamic speech imaging. Without loss of generality, let us assume the ***(k**, t)*-space data points fall on a Cartesian grid at discrete time points. The sampled data points can be expressed as

$$d(m\Delta k_x, n\Delta k_y, t) = \int \rho(x,y,t) e^{-i2\pi(m\Delta k_x x + n\Delta k_y y)} dxdy, \tag{3.3}$$

where Δk_x and Δk_y denote sampling spacing in the frequency encoding and phase encoding directions respectively, t denotes the time instant for discrete sampling. If we further assume that the targeted image function, $\rho(x,y,t)$, has bounded spatial support in the ***(x**, f)*-space,

$$\rho(x,y,t) = 0 \text{ for } |x| > FOV_x,\ |y| > FOV_y, \tag{3.4}$$

where FOV_x and FOV_y denote the field-of-view in each spatial direction. The Nyquist sampling theorem asserts that the ***k***-space intervals Δk_x and Δk_y must satisfy the following requirement in order to avoid spatial aliasing in the reconstructed image function $\rho(x,y,t)$ [48],

$$|\Delta k_x| \le \frac{1}{FOV_x},\ |\Delta k_y| \le \frac{1}{FOV_y}, \tag{3.5}$$

where the reconstructed image has a spatial resolution defined by the following inequalities,

$$|\Delta x| \ge \frac{1}{|\Delta k_x| N_x},\ |\Delta y| \ge \frac{1}{|\Delta k_y| N_y}, \tag{3.6}$$

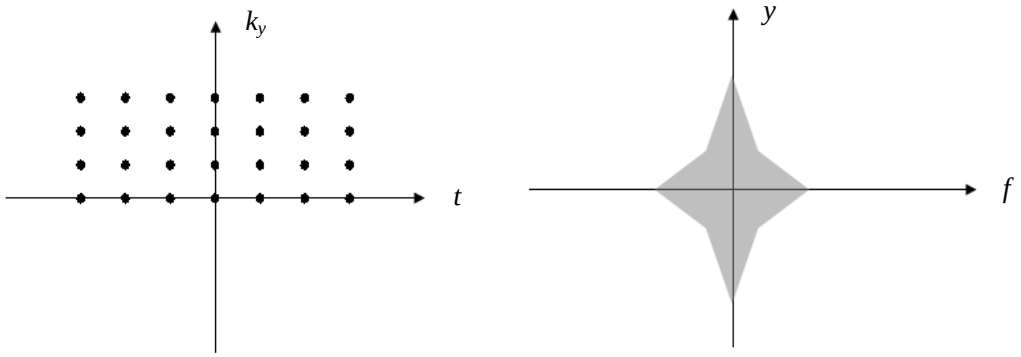
where N_x denotes number of samples in the horizontal direction and N_y denotes the number of samples in the vertical direction. Let T_R denote the sampling time it takes to collect all samples from a readout line in the ***k***-space and if Cartesian sampling is assumed in this case, the total data acquisition time *T* is given by

$$T = N_y T_R. \tag{3.7}$$

Given the limited MR imaging speed, if the temporal Nyquist rate were to be satisfied, the tem-

poral resolution is usually increased by reducing the number of phase encoding lines. This is because the value of T_R is usually constrained by both physical and physiological considerations. However, according to inequalities (3.6), reducing the number of phase encoding lines gives rise to worse spatial resolution. Therefore, it is unrealistic to achieve temporal Nyquist rate sampling and spatial Nyquist rate sampling at the same time. This is the intrinsic compromise that limits Nyquist rate sampling in the ***(k**, t)*-space.

The above compromise can be better understood in the ***(k**, t)*-space and its conjugate space, the ***(x**, f)*-space. Figure 3.1(a) depicts an ideal instantaneous sampling pattern for an exemplary signal in the ***(k**, t)*-space. If this signal is acquired with the ideal sampling pattern, in the ***(x**, f)*-space we have a un-aliased bounded support for the signal, as can be seen in Fig. 3.1(b). However, in practice the signal is usually sampled in a time-sequential manner, i.e., only one phase encoding line can be sampled within any time instant. The corresponding time-sequential sampling pattern, as depicted in Fig. 3.2(a), can be regarded as a temporally undersampled pattern of the ideal sampling pattern. This temporal under-sampling reduces sampling data points in the ***(k**, t)*-space and hence increases temporal resolution.



(a) (b)

Figure 3.1: (a) The ideal instantaneous sampling pattern in the ***(k**, t)*-space. Each black dot denotes a sampled phase encoding line. (b) The bounded spatial-spectral support for a typical dynamic signal.

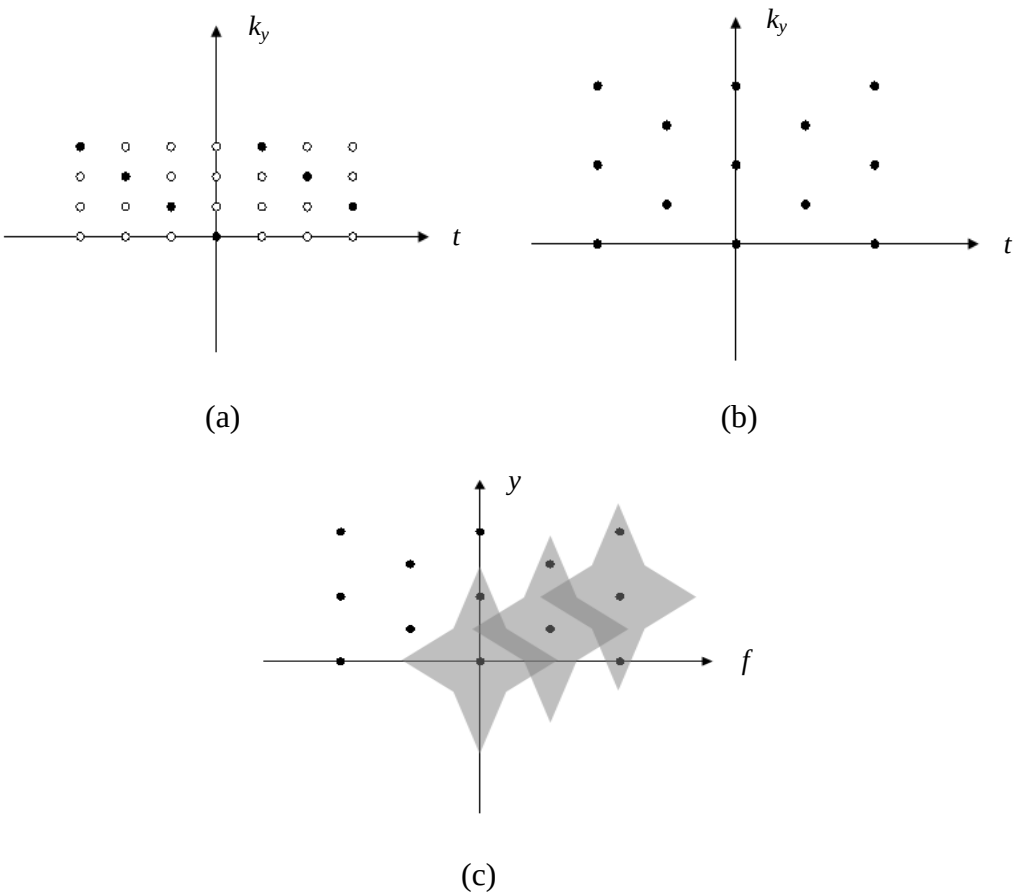


Figure 3.2: (a) The time-sequential sampling pattern in the ***(k**, t)*-space. Black dots denote sampled phase encoding lines, while white dots denote phase encoding lines that are not sampled. (b) The point spread function in the ***(x**, f)*-space corresponding to the sampling pattern in (a). (c) Replicas of the original spectra overlap in the ***(x**, f)*-space. Overlapping in the ***(x**, f)*-space introduce aliasing.

According to sampling and interpolation theory, since the ***(k**, t)*-space and the ***(x**, f)*-space are related by the Fourier transform, temporal under-sampling creates replicates in the ***(x**, f)*-space. These replicates are arranged by the point spread function in the ***(x**, f)*-space, which is the spatial Fourier transform of the ***(k**, t)*-space sampling pattern. The point spread function in the ***(x**, f)*-space is depicted in Fig. 3.2 (b). Let us assume the ***(k**, t)*-space is highly undersampled in time. In this case, although a high reduction factor is achieved in the overall sampling time, the spacing between point spread function corresponding to this ***(k**, t)*-sampling pattern will be “squeezed”

on the frequency direction. Since the $(\boldsymbol{x}, f)$-spectrum can be regarded as the convolution of the original spectra with a point spread function, replicates of $(\boldsymbol{x}, f)$-spectra will be densely allocated according to the point spread function. As a result, these replicates will overlap in the $(\boldsymbol{x}, f)$-space, as is depicted in Fig. 3.2 (c). This overlap will in turn cause irresolvable motion-related artifacts in the reconstructed image. Under the $(\boldsymbol{k}, t)$-space and $(\boldsymbol{x}, f)$-space framework, overlapping $(\boldsymbol{x}, f)$-spectrum is the fundamental reason that prevents high temporal resolution and high spatial resolution from being achieved simultaneously.

The above analysis also suggests the significance of increasing the data acquisition speed. The intrinsic compromise between the temporal resolution and spatial resolution is, by its nature, caused by the limited MR imaging speed. If the data acquisition speed can be effectively boosted, both temporal resolution and spatial resolution can be increased and sharper temporal variations can be observed within a same period of time since data acquisition speed, spatiotemporal resolution and speech dynamics are interdependent parameters. Improvement in these parameters is of great value to dynamic speech and cardiac imaging research. Furthermore, reduction in data acquisition time accelerates the overall MR scanning procedure and is a practical concern in reducing long MR examination for clinical purposes. Given these benefits, it is natural to increase the imaging speed by resorting to two general approaches: First, to sample the $(\boldsymbol{k}, t)$-space with higher sampling rate. Second, to sample less $(\boldsymbol{k}, t)$-space within the same period of time. Compared with the first approach, the second approach is apparently more difficult since it requires advanced $(\boldsymbol{k}, t)$-space models to sparsely sample and reconstruct the acquired data.

3.1.2 Existing sparse sampling strategies in (k, t)-space

Three fundamental approaches exist to accelerate data acquisition in dynamic MR imaging: the pulse sequence-based approach [37 - 39], the parallel imaging-based approach [40 - 42] and the signal processing-based approach [49 - 59]. The first two approaches have gone through decades of development and have been thoroughly discussed in relevant literatures [37 - 42]. However,

the signal processing-based approach not only enables accelerated imaging, but can also incorporate fast pulse sequences and advanced hardware systems to yield even higher imaging speed. Therefore, in this thesis we will briefly introduce the sequence-based and hardware-based approaches and focus our discussion on the signal processing-based approach.

The sequence-based methods aim to sweep through the $(\boldsymbol{k}, t)$-space at higher imaging speed. Specifically, imaging speed is accelerated by combining fast pulse sequences, efficient trajectories and high-performance gradient systems to traverse the $(\boldsymbol{k}, t)$-space faster. Examples of fast pulse sequences include FLASH [37], FISP [38], EPI [39] and their variants. These pulse sequences can be implemented with various $\boldsymbol{k}$-space trajectories such as the Cartesian trajectories, the radial trajectories and the spiral trajectories to improve imaging speed. After decades of development, the sequence-based methods have already provided a variety of fast imaging options. However, the potential of these methods is mainly constrained by the energy deposit due to RF exposure and potential stimulation to the human peripheral nerve system.

The parallel imaging-based methods improve temporal resolution by exploiting the physical potential of specialized devices, such as multichannel receiver coils. Parallel imaging is a representative technique that falls into this category. Examples of typical parallel imaging techniques include SENSE [40], SMASH [41], GRAPPA [42] and their variants. Although these techniques may differ in their specific strategies to combine multichannel data, invariably they distribute the phase encoding portions of the field of view among an array of receiver coils and carry out these tasks in parallel. Sensitivity information of each coil allows the acquired data from multiple receiver coils to be later combined into a full field of view image [40]. However, the ability of parallel imaging techniques to accelerate dynamic imaging is mainly compromised by the SNR penalty [49]. As has been discovered in some literatures, SNR is proportional to the square root of total MR scanning time [49]. With reduced scanning time, the SNR in parallel imaging is consequently lower than conventional data acquisition. SNR may be further reduced by the specific image reconstruction technique used, such as SENSE and SMASH [49]. Despite parallel imaging techniques, other hardware-based sampling techniques also have their intrinsic limits.

The signal processing-based methods accelerate the imaging speed by taking advantage of the spatiotemporal properties of the image function. Specifically, spatiotemporal correlation allows sparse sampling of the $(\boldsymbol{k}, t)$-space without significantly losing signal dynamic. Sparse sampling can be usually performed either in time, in $\boldsymbol{k}$-space or in both in time and $\boldsymbol{k}$-space. Since the sparsely sampled data often lead to aliasing, an effective imaging method usually needs to employ additional constraints to reconstruct signal dynamics. By incorporating different constraints to regularize reconstruction, images recovered from sparse sampling can have various properties.

A variety of signal processing-based methods have been proposed to accelerate dynamic imaging. One way to categorize these methods is through the constraints that are imposed in image reconstruction. In this way, most of the signal processing-based methods can be placed into one of four categories: (1) Methods that rely on the packing of bounded spatial-spectral support. Examples of these methods include UNFOLD [50], TSENSE [51], PARADIGM [52] and PARADISE [53]. (2) Methods that rely on the spatiotemporal correlation in acquired data. Examples of these methods include *k-t* SENSE [54] and *k-t* BLAST [55]. (3) Methods that rely on signal sparsity in a transformed domain. Examples of these methods include *k-t* FOCUSS [56] and *k-t* SPARSE [57]. (4) Methods that rely on partial separability of the dynamic signal. Examples of these methods include PS [58] and PS SPARSE [59]. While there have been other alternative schemes to sample and reconstruct the $(\boldsymbol{k}, t)$-space data, this thesis does not aim to provide an exhaustive review of all the methods.

3.1.2.1 UNFOLD and PARADIGM

UNFOLD (un-aliasing by Fourier-encoding the overlaps using the temporal dimension) and PARADIGM (patient-adapted reconstruction and acquisition dynamic imaging method) are two representative imaging schemes that are based on the packability of the bounded spatial-spectral support. Let us start from UNFOLD to explain the mechanism of these schemes.

UNFOLD is based on three underlying assumptions [60]. Firstly, a dynamic signal is as-

sumed to be comprised of two portions, a dynamic portion and a static portion [60]. Secondly, both the dynamic portion and the static portion are assumed to be support limited in the *(**x**, f)*-space [60]. Thirdly, the "effective" support in the *(**x**, f)*-space should be packable [60]. Figure 3.3 depicts an example of a typical *(**x**, f)*-support consisting of two rectangles. As can be seen, the "fat and tall" rectangle represents the dynamic portion that has larger temporal frequency span but less spatial coverage [60]. On the contrary, the "thin and flat" rectangle represents the static portion that has less temporal frequency bandwidth but more spatial coverage [60]. The signal intensity outside this "cross-shape" support is usually assumed to be zero.

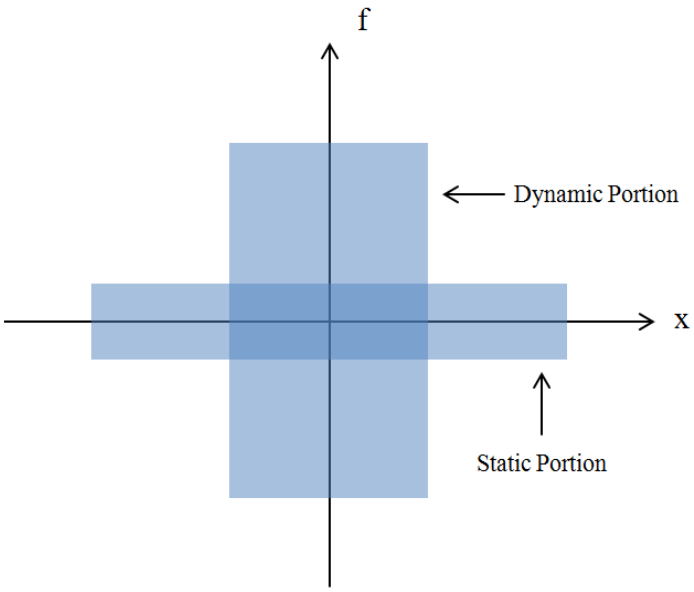


Figure 3.3: UNFOLD assumes "cross-shape" support in the *(**x**, f)*-space.

With these three assumptions, the data acquisition scheme in UNFOLD is designed in a way that both the dynamic portion and the static portion are allowed to reside in the same temporal frequency interval but are modulated with a distinctive phase [50]. The original UNFOLD algorithm introduces spatial shift in *(**k**, t)* sampling pattern to realize this goal [50]. Suppose a shift Δk_s is applied to the sampling pattern in the phase encoding direction,

$$d_{acq}(k_y, t) = d(k_y, t) \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \delta(k_y - p\Delta k_y - \Delta k_s, t - q\Delta t), \quad (3.8)$$

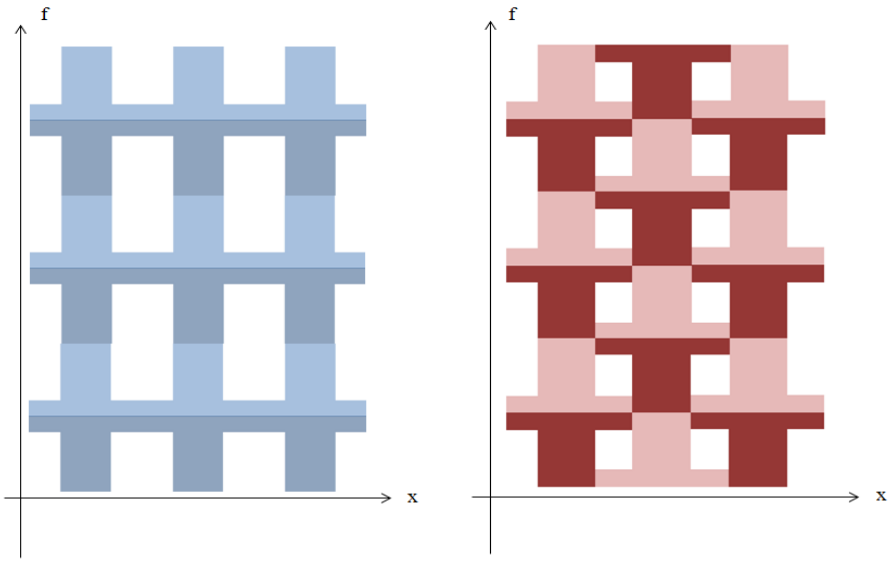
where Δk_y denotes the spatial sampling interval in the phase encoding direction, Δt denotes the temporal sampling interval, $d(k_y, t)$ denotes the desired image function in the ***(k**, t)*-space and $d_{acq}(k_y, t)$ denotes the acquired data. If the inverse spatial Fourier transform is applied to both sides of equation (3.8), the following relation can be easily obtained,

$$\begin{aligned}\rho_{acq}(y,t) &= \rho(y,t) * \mathcal{F}^{-1}\{\textstyle\sum_{p=-\infty}^{\infty}\sum_{q=-\infty}^{\infty}\delta(k_y - p\Delta k_y - \Delta k_s, t - q\Delta t)\},\\ &= \rho(y,t) * \left[\frac{1}{\Delta k_y} e^{i2\pi y \Delta k_s} \textstyle\sum_{q=-\infty}^{\infty}\sum_{p=-\infty}^{\infty}\delta\left(y - \frac{p}{\Delta k_y}, t - q\Delta t\right)\right],\\ &= \frac{1}{\Delta k_y}\textstyle\sum_{q=-\infty}^{\infty}\sum_{p=-\infty}^{\infty} e^{i2\pi\frac{p}{\Delta k_y}\Delta k_s}\rho\left(y - \frac{p}{\Delta k_y}, t - q\Delta t\right). \end{aligned} \quad (3.9)$$

The right-hand side of equation (3.9) denotes summation over the replicas of the desired image function modulated by a phase term of exp $(i2\pi p\Delta k_s/\Delta k_y)$. From this expression, it is obvious that the desired image function can be separated by applying a temporal filter if there is no overlap in the *(**x**, f)*-space [61].

The data acquisition scheme in UNFOLD can also be considered as a lattice sampling scheme in the *(**k**, t)*-space [61]. In the lattice sampling scheme presented in [61], the corresponding "cross-shaped" *(**x**, f)*-support can be packed more densely in a way that minimal "wasted space" is left. From this perspective, manipulation of the phase term in equation (3.9) is by its nature seeking a tighter way to embed signal support in the *(**x**, f)*-space [61]. This can be illustrated with Fig. 3.4. Figure 3.4(a) depicts the conventional sampling pattern in the *(**x**, f)*-space. In this pattern, replicates of the *(**x**, f)*-spectrum are arranged on a Cartesian grid. If this pattern is used for sampling, a lot of empty *(**x**, f)*-space is created and thus burdens the sampling requirement in the *(**k**, t)*-space. However, Fig. 3.4 (b) depicts the UNFOLD sampling pattern in the *(**x**, f)*-space. In this pattern, replicates of the *(**x**, f)*-spectrum are arranged on a sheared lattice. Compared to the conventional sampling pattern in Fig. 3.4(a), the UNFOLD pattern has a tighter spatial arrangement in the *(**x**, f)*-space, which is helpful for reducing the sampling requirement in the *(**k**, t)*-space [61]. These results demonstrate how UNFOLD accelerates *(**k**, t)*-space sampling with

efficient spectral packing.



(a) (b)

Figure 3.4: (a) Conventional *(**x**, f)*-sampling pattern. (b) UNFOLD *(**x**, f)*-sampling pattern.

UNFOLD provides a straightforward reconstruction technique for the sparsely sampled data in the *(**k**, t)*-space [50]. Its significance not only lies in increasing the temporal resolution of the reconstructed image, but also sheds light upon viewing *(**k**, t)*-space sampling as a lattice sampling scheme [61]. Also, the UNFOLD method can be combined with parallel imaging techniques to yield further acceleration. However, the acceleration factor of UNFOLD is mainly restricted by its assumption on the "cross-shape" *(**x**, f)*-support, which prevents UNFOLD from packing support in a denser way.

PARADIGM is different from UNFOLD in its assumption on the *(**x**, f)*-support. UNFOLD assumes a "cross shape" *(**x**, f)*-support for all the objects to be imaged. In reality, however, it is unwise to make such an assumption since the support information is patient-dependent and application dependent [52]. If wrong assumptions about the *(**x**, f)*-support are made, model mismatch in reconstruction is likely to cause aliasing. Instead of assuming a fixed *(**x**, f)*-support for every object, PARADIGM adopts a patient-specific *(**x**, f)*-support and offers patient-adaptive reconstructions based on this support [52]. Specifically, PARADIGM makes the following three

assumptions: Firstly, the *(**x**, f)*-support with non-negligible energy is band-limited and packable [52]. Secondly, the *(**x**, f)*-support is assumed to possess a harmonically-related multiband structure [52]. This multiband structure comes naturally as the result of periodic or quasi-periodic motion in the region of interest [52]. Periodicity in dynamic motion suggests that the *(**x**, f)*-support does not cover a broad temporal frequency spectrum; rather, only a few locations on the temporal frequency axis are covered. In addition, the width of the *(**x**, f)*-bands depends on the temporal variation of the specific dynamic motion being imaged. Greater motion variability usually accompanies wider width of the multiband structure and *vice versa*. Thirdly, reconstruction of a dynamic image sequence can be obtained from simply filtering the acquired data with a band-pass filter [52]. These three assumptions allow for a more precise determination of *(**x**, f)*-support and result in a tighter lattice sampling scheme than that of UNFOLD.

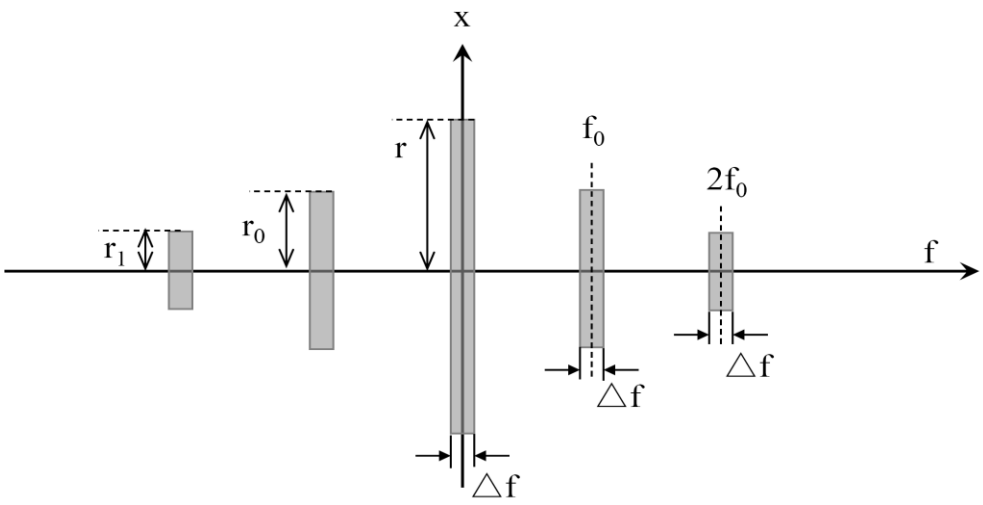


Figure 3.5: The PARADIGM *(**x**, f)* –support.

The PARADIGM scheme can be decomposed into three major steps: the preparation step, the data acquisition step and the image reconstruction step [52]. In the preparation step, a navigator signal is acquired on the object to be imaged in order to determine the ***k***-space region of interest K as well as the *(**x**, f)*-support S. K is usually determined as the union of all ***k***-space region of interest estimated from the navigator signal [52]. Similarly, S is usually determined as the union

of all ***(x**, f)*-support estimated from the navigator signal [52],

$$S = \bigcup_{(x,f)} supp_{(x,f)}\{\mathcal{F}_t[\rho(\boldsymbol{x}, t)]\} \ , \tag{3.10}$$

where $supp_{(x,f)}$ denotes the ***(x**, f)*-support and $\mathcal{F}_t$ denotes the temporal Fourier transform matrix. If the object to be imaged in PARADIGM has periodic or quasi-periodic motion [52], according to the Fourier transform properties, S displays a multiband structure in the ***(x**, f)*-space, as is depicted in Fig. 3.5.

With prior knowledge about the ***(x**, f)*-support, the sampling strategy in the data acquisition step aims to design an optimized trajectory that minimize the overall data acquisition time [52]. The design of an optimized sampling scheme is based on two basic assumptions. Firstly, the ***(x**, f)*-support information, especially S and K, can be effectively determined from the navigator signal. Secondly, a lattice sampling scheme is used to acquire data due to the physical and physiological limitations during data acquisition [52]. These assumptions guarantee that the sampling scheme designed in the data acquisition step is patient-adaptive and realistic for hardware implementation. Specifically, lattice sampling aims to optimize the sampling period [52],

$$T_{opt}(S,\ K) = T(\Lambda_{opt}(S,\ K)), \tag{3.11}$$

where T denotes the sampling period, T_{opt} denotes the optimized sampling period, Λ_{opt} denotes the optimized lattice on which data are acquired, which can be determined by ,

$$\Lambda_{opt}(S,\ K) = \underset{\Lambda^* \in P(S)}{\arg\max}\, T(\Lambda), \tag{3.12}$$

where Λ denotes the sampling lattice, $P(S)$ denotes the ensemble of all lattices that span S [52]. In practice, the optimized sampling lattice can be determined by imposing some geometric constraints [52].

Given the acquired data, image reconstruction of $\rho(\boldsymbol{x}, t)$ in PARADIGM is straightforward. Specifically, reconstruction can be performed in two steps. Firstly, the acquired data $d(\Lambda)$ is processed with a multidimensional band-pass filter, which is equivalent to convolving the ac-

quired data with an interpolation kernel $\Phi(\boldsymbol{k}, t)$ [52]. Mathematically, this can be expressed as,

$$\hat{d}(\boldsymbol{k}, t) = \sum_{\boldsymbol{\gamma}\epsilon\Lambda} d(\Lambda)\, \Phi\left[\binom{\boldsymbol{k}}{t} - \boldsymbol{\gamma}\right], \tag{3.13}$$

where convolution in equation (3.13) is performed in both $\boldsymbol{k}$ and t [52]. Secondly, the desired image sequence can be calculated from the Fourier transform of the filtered data. Since the above equation involves summation within only finite steps, the reconstruction error can be easily controlled by changing a series of parameters, such as the shape of interpolation kernel Φ and the sampling period T [52]. Previous research has also given a series of theoretical bounds for various image quality metrics and image acquisition parameters [52].

Compared to UNFOLD, PARADIGM offers a more adaptive imaging scheme since its *(**x**, f)*-support is determined from the patient-specific navigator signal rather than a predetermined geometric shape [52]. In addition, a higher level of *(**x**, f)*-support packability is achieved in PARADIGM since the harmonically-related multiband structure makes better use of the *(**x**, f)*-space than the "cross shape" support in UNFOL [52]. Greater packability in PARADIGM further reduces the temporal and spatial sampling requirements in the *(**k**, t)*-space compared to UNFOLD. However, feasibility of PARADIGM lies in the assumption on the multiband *(**x**, f)*-support. If the signal dynamics in the region of interest severely violate periodicity, the performance of PARADIGM may be compromised since the harmonic band structure no longer holds.

3.1.2.2 *k-t* BLAST and *k-t* SENSE

k-t BLAST (*k-t* broad-use linear acquisition speedup technique) and *k-t* SENSE (*k-t* sensitivity encoding) are two representative imaging schemes that take advantage of the spatiotemporal correlations to allow sparse sampling and reconstruction from sparsely sampled data. *k-t* BLAST and *k-t* SENSE share the same imaging principle and imaging scheme but can be applied to different dynamic imaging situations. Specifically, *k-t* BLAST is used with a single receiver coil, while *k-t* SENSE is used with multiple receiver coils.

If a dynamic imaging signal is very sparsely sampled in the ***(k**, t)*-space, according to basic Fourier transform relationship, the corresponding aliased ***(x**, f)*-spectrum can be regarded as an ensemble of superimposed original signal spectrum arranged according to a point spread function [63]. The purpose of *k-t* BLAST and *k-t* SENSE, therefore, is to un-alias each voxel by redistributing its value back to each contributing voxel before the temporal Fourier transform is applied to obtain the original time sequence [63]. Without loss of generality, let us start from *k-t* SENSE to explore how aliasing in ***(x**, f)*-space can be resettled. Mathematically, this is realized by modeling the un-aliasing process as a minimum norm problem, where estimated value of each contributing voxel is determined by the weighted minimum norm solution [63]. If we further assume that the imaging system has time invariant coil sensitivity, un-aliasing in *k-t* BLAST can be briefly expressed as the following equation,

$$\boldsymbol{\rho} = \boldsymbol{C}(\mathbf{1}\,\boldsymbol{C})^{+}\boldsymbol{\rho}_{al}, \tag{3.14}$$

where $\boldsymbol{\rho}$ denotes the desired un-aliased image in ***(x**, f)*-space, $\boldsymbol{C}$ denotes a diagonal matrix with its diagonal entries representing an estimate of signal magnitudes of the contributing voxel, $\mathbf{1}$ denotes a row vector with all entries being 1 and $\boldsymbol{\rho}_{al}$ denotes the aliased image in the ***(x**, f)*-space [63]. Solution in equation (3.14) can be obtained by expanding the Moore-Penrose pseudo inverse in equation (3.14) [63],

$$\boldsymbol{\rho} = \boldsymbol{C}^2\mathbf{1}^H(\mathbf{1}\,\boldsymbol{C}^2\mathbf{1}^H)^{-1}\boldsymbol{\rho}_{al}. \tag{3.15}$$

With simple mathematical manipulation, equation (3.15) can be further expanded as [63],

$$\boldsymbol{\rho} = \frac{\left[\prod_{n=1}^{N}(c_n)^2\right]^H}{\sum_{n=1}^{N}(c_n)^2}\boldsymbol{\rho}_{al}\,, \tag{3.16}$$

where c_n denotes the diagonal entries of $\boldsymbol{C}$ and *N* denotes the number of voxels. Equation (3.16) suggests that $\boldsymbol{\rho}$ can be recovered from $\boldsymbol{\rho}_{al}$ by reallocating the aliased voxel values based on estimates of relative signal power of each contributing voxel [63]. From an imaging perspective, the relative signal power in equation (3.16) determines the region in which signal variation is more likely to concentrate in the ***(x**, f)*-space and serves as a reference for realloca-

tion of voxel values [63].

From equation (3.15) and equation (3.16), it is obvious that the feasibility of *k-t* BLAST and *k-t* SENSE lies in an accurate estimation of the magnitudes of contributing voxels [63]. In *k-t* BLAST and *k-t* SENSE, a training data set is introduced to obtain an estimate of these voxel magnitudes before the original image is un-aliased. Specifically, both *k-t* BLAST and *k-t* SENSE adopt a composite data acquisition scheme that is made up of two major steps, the acquisition of a training data and the acquisition of an imaging data [63]. The training data set is used to provide estimate of voxel magnitudes and the imaging data set is used to recover the targeted image sequence [63]. An example of this composite imaging scheme is depicted in Fig. 3.6.

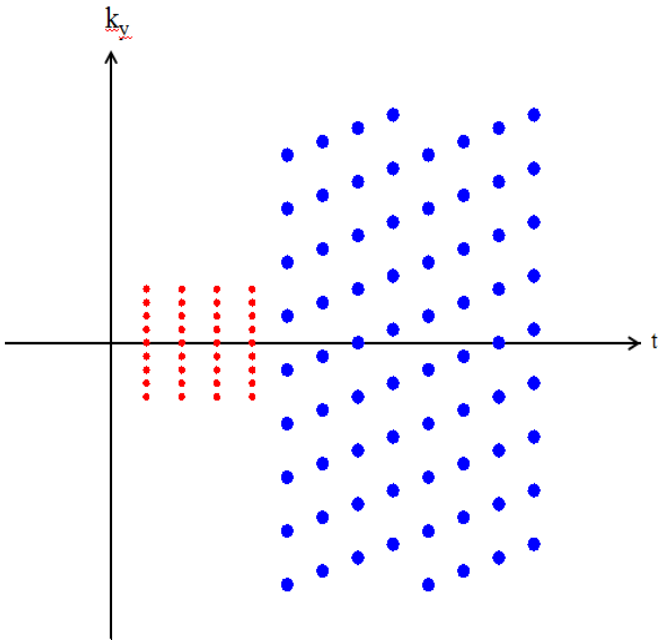


Figure 3.6: The ***(k**, t)*-sampling pattern in *k-t* BLAST and *k-t* SENSE. Red dots denote the training data set and blue dots denote the imaging data set.

From Fig. 3.6, it is obvious that the training data set is acquired in high temporal resolution and low spatial resolution. On the contrary, the imaging data set is acquired in high spatial resolution and low temporal resolution. In the training data set, the resultant low spatial resolution image is usually used to determine the diagonal entries of the weighting matrix $\boldsymbol{C}$, which is later used to facilitate the reconstruction of $\boldsymbol{\rho}$ according to equation (3.16). In practice, *k-t* BLAST and *k-t*

SENSE have further consideration: $\boldsymbol{\rho}$ is regarded as being fluctuating around a baseline ue $\boldsymbol{\rho}_{bl}$ [63], which is also incorporated into equation (3.15) to assist image reconstruction. Given $\boldsymbol{C}$ and $\boldsymbol{\rho}_{bl}$, *k-t* BLAST modifies equation (3.15) and straightforwardly carry out reconstruction according to the following equation [63],

$$\boldsymbol{\rho} = \boldsymbol{\rho}_{bl} + \boldsymbol{C}^2\mathbf{1}^H(\mathbf{1}\boldsymbol{C}^2\mathbf{1}^H + \psi)^{-1}(\boldsymbol{\rho}_{al} - \mathbf{1}\boldsymbol{\rho}_{bl}), \tag{3.17}$$

where ψ denotes the noise variance. For *k-t* SENSE, since sensitivity information of each receiver soil can be obtained, the above equation can be generalized to the following equation [63],

$$\boldsymbol{\rho} = \boldsymbol{\rho}_{bl} + \boldsymbol{C}^2\boldsymbol{S}^H(\boldsymbol{S}\boldsymbol{C}^2\boldsymbol{S}^H + \boldsymbol{\Psi})^{-1}(\boldsymbol{\rho}_{al} - \boldsymbol{S}\boldsymbol{\rho}_{bl}). \tag{3.18}$$

Replacing the row vector $\mathbf{1}$ in *k-t* BLAST with $\boldsymbol{S}$ in *k-t* SENSE brings twofold benefits. On the one hand, an underdetermined problem in *k-t* BLAST can be transformed into an overdetermined problem in *k-t* SENSE under some specific conditions [63]. On the other hand, since the obtained training data has limited spatial resolution, the sensitivity information can be used to provide extra information to assist reconstruction. In this manner, *k-t* SENSE excels *k-t* BLAST in a way that more information can be combined for reconstruction [63]. From the above description of both *k-t* SENSE and *k-t* BLAST, it is not difficult to find their shared dependence on one fundamental assumption: the motion pattern learned from training data should be representative of the motion occurred in the imaging data set [63]. As long as this motion pattern remains the same for both data sets, reconstruction according to equation (3.17) and equation (3.18) can be used to represent genuine signal dynamics [63]. However, if the motion patterns obtained from the two stages do not match, some model-induced misregistration may degrade final reconstruction.

3.1.2. 3 *k-t* FOCUSS and *k-t* SPARSE

k-t SPARSE and *k-t* FOCUSS are two representative compressed sensing based imaging schemes that take advantage of signal sparsity in transformed domains [56, 57]. Specifically *k-t* SPARSE makes use of signal sparsity in the wavelet-frequency domain and *k-t* FOCUSS makes use of

signal sparsity in the spatial-temporal frequency domain. Although both imaging schemes may vary in their specific algorithm and implementation, they share their common origin in the compressed sensing theory. Before we discuss the details about both schemes, let us first briefly review the compressed sensing theory.

The compressed sensing theory is deeply rooted in the compressibility or sparsity in natural images [56]. Sparsity of an acquired image suggests that a sparsifying transform can be applied to convert that image into a certain transform domain where only a few non-zero coefficients exist [64]. Given the compressibility of an image or sparsity of the targeted signal, the compressed sensing scheme aims to sample a compressed amount of data, instead of the amount of data indicated by the Nyquist criterion, and recover the undersampled data through sparsity-promoting algorithms [64]. Compared with conventional approaches that acquire a large amount of data, compressed sensing theories depend heavily on the inherent correlation in the signal of interest to significantly reduce sampling requirements [64].

Mathematically, the compressed sensing theory aims to obtain a sparse solution for the following equation [64],

$$\mathbf{d} = \mathbf{A}\boldsymbol{\rho} + \boldsymbol{\epsilon}\,, \tag{3.19}$$

where $\mathbf{d}$ denotes a length-M measured data, $\mathbf{A}$ denotes an M $\times$ N encoding matrix, $\boldsymbol{\rho}$ denotes the desired image of interest, and $\boldsymbol{\epsilon}$ denotes a noise vector with a length of *M*. Since $\mathbf{A}$ is usually regarded as a matrix with full row rank, equation (3.19) is an underdetermined system with a large number of candidate solutions. However, compressed sensing theory guarantees a unique solution to the above inverse problem by making two assumptions. Firstly, the desired solution is assumed to be sparse or highly compressible [64],

$$\hat{\boldsymbol{\rho}} = \boldsymbol{\Phi}\boldsymbol{\rho}\ , \tag{3.20}$$

where $\boldsymbol{\Phi}$ is a sparsifying transform matrix and $\hat{\boldsymbol{\rho}}$ is a sparse representation of $\boldsymbol{\rho}$ in the transformed domain. Secondly, matrix $\mathbf{A}$ is assumed to conform to some specific mathematical conditions [64]. Under these two assumptions, the compressed sensing theory eliminates a large

number of candidate solutions to the sparsest solution that maximizes the number of non-zero coefficients of sparse representation while remaining consistent with the measured data **d** [64]. This process can be expressed as [64]

$$\begin{aligned} &\text{maximize } H_s(\boldsymbol{\rho}) \\ &\text{subject to } \|\boldsymbol{A\rho} - \boldsymbol{d}\|_{\mathbf{2}} < \epsilon \end{aligned}, \tag{3.21}$$

where $H_s(\cdot)$ denotes a function that evaluates the sparsity of $\boldsymbol{\rho}$, $\|\cdot\|_{\mathbf{2}}$ denotes the ℓ_2 norm, and ϵ denotes the noise level. A natural choice of $H_s(\cdot)$ is the ℓ_0 norm since the ℓ_0 norm, by its definition, directly sums all the non-zero entries of $\boldsymbol{\rho}$. However, the use of the ℓ_0 norm is not realistic in practice because solving the ℓ_0 norm minimization problem often leads to an NP hard problem [64], i.e., the problem cannot be settled within polynomial time. Therefore, it is usually preferable to resort to some surrogate options that serve the same purpose as the ℓ_0 norm but are solvable in practice. Although many options exist as the surrogate functions for the ℓ_0 norm [65, 66, 67], in this thesis we focus only on the ℓ_1 norm. The ℓ_1 norm and the ℓ_1 norm minimization method are chosen because the ℓ_1 norm is found to be the tightest convex relaxation to the ℓ_0 norm [64]. Under some mathematical conditions, the solution to the ℓ_1 norm minimization problem can be regarded as a satisfying approximation to the ℓ_0 norm solution [68]. In this manner, the original optimization problem can be rewritten as [68],

$$\begin{aligned} &\text{minimize } \|\boldsymbol{\rho}\|_1 \\ &\text{subject to } \|\boldsymbol{A\rho} - \boldsymbol{d}\|_{\mathbf{2}} < \epsilon \end{aligned}, \tag{3.22}$$

where $\|\cdot\|_{\mathbf{1}}$ denotes the ℓ_1 norm. With formulation (3.22), the compressed sensing theories assert that the reconstruction error for $\boldsymbol{\rho}$ can be constrained within a theoretical performance bound when the restricted isometric property (RIP) is satisfied [69]. RIP starts by defining δ_s as the minimum coefficient for equation (3.23) to be satisfied,

$$(1 - \delta_s)\|\boldsymbol{\rho}\|_2^2 \leq \|\boldsymbol{A\rho}\|_2^2 \leq (1 + \delta_s)\|\boldsymbol{\rho}\|_2^2. \tag{3.23}$$

where $\boldsymbol{\rho}$ is an s-sparse signal. Equation (3.23) suggests that better reconstruction of $\boldsymbol{\rho}$ can be obtained with a smaller δ_s and previous research has proven that, if noise-free sampling is satisfied in practice, equation (3.23) can guarantee perfect reconstruction of $\boldsymbol{\rho}$ when $\delta_{2s} < \sqrt{2} -$

1 [69]. Even when samples of $\boldsymbol{\rho}$ are contaminated by noise, equation (3.23) defines a theoretical performance bound for the reconstruction error $\boldsymbol{\eta}$ [70],

$$\|\boldsymbol{\eta}\|_2 \leq B_0 s^{-\frac{1}{2}} \|\boldsymbol{\rho} - \boldsymbol{\rho}_s\|_1 + B_1 \epsilon, \tag{3.24}$$

where $\boldsymbol{\eta}$ denotes the reconstruction error between the true image $\boldsymbol{\rho}$ and the one recovered from the ℓ_1 norm minimization method, $\boldsymbol{\rho}_s$ denotes the optimal s-sparse approximation of $\boldsymbol{\rho}$, B_0 and B_1 are functions that depend on δ_{2s} [70]. This performance bound guarantees perfect recovery of the original image $\boldsymbol{\rho}$ from the sparsely sampled data under some mathematical constraints [70]. Its existence sheds light upon many compressed-sensing based imaging schemes that are tailored for MR experiments. The *k-t* SPARSE is one representative imaging scheme.

k-t SPARSE can be regarded as a special example of the compressed sensing-based imaging schemes. In *k-t* SPARSE, the wavelet-temporal-frequency space is employed to assist reconstruction of the dynamic image sequence [56]. Specifically, image reconstruction in *k-t* SPARSE can be mathematically expressed as,

$$\begin{aligned} &\text{minimize } \|\boldsymbol{W}\boldsymbol{\rho}\|_1 \\ &\text{subject to } \|\boldsymbol{A}\boldsymbol{\rho} - \boldsymbol{d}\|_2 < \epsilon \text{ '} \end{aligned} \tag{3.25}$$

where $\boldsymbol{W}$ represents a sparsifying transform in both the wavelet domain and the temporal frequency domain [56]. These two domains are combined in a way that wavelet transform can be used to sparsify spatial variations of $\boldsymbol{\rho}$ and coefficients from Fourier transform are used to sparsify periodic or sometimes quasi-periodic temporal behaviors of $\boldsymbol{\rho}$ [56]. Compared with other imaging methods that are based on the compressed sensing theory, *k-t* SPARSE is a straightforward extension in the wavelet-temporal-frequency domain and can be refined by adopting tailored regularizations, advanced sampling schemes and specialized reconstruction methods to serve specific imaging purposes [71].

Unlike *k-t* SPARSE, *k-t* FOCUSS takes advantage of signal sparsity in the *(**x**, f)*-space [57]. Despite its sparsity-promoting nature, *k-t* FOCUSS has its root in many other sources. Perhaps its closest origin, as its name suggests, is the original FOCUSS algorithm [72]. The FOCUSS algo-

rithm aims to achieve sparse solution through successive quadratic optimization [72]. This strategy is adopted in *k-t* FOCUSS and is further developed to yield sparse *(**x**, f)*-support through multiple iterations. Despite its similarity with the FOCUSS algorithm, *k-t* FOCUSS is also closely related to *k-t* BLAST, *k-t* SENSE and the compressed sensing theory [54 - 56]. However, *k-t* FOCUSS is different from these theories by reformulating the sparse sampling strategies of *k-t* BLAST and *k-t* SENSE in another framework [72 - 73]. Compared with *k-t* BLAST or *k-t* SENSE, *k-t* FOCUSS allows for more iterations to redefine signal support in the *(**x**, f)*-space and adopts specific updating strategies to yield higher spatiotemporal resolution [73]. *k-t* FOCUSS also incorporates compressed sensing theories to guide sparse sampling in *(**k**, t)*-space and proves that the FOCUSS solution is in fact equivalent to ℓ_1 norm minimization under certain mathematical conditions [57]. In this manner, *k-t* FOCUSS bridges the connection between *k-t* BLAST, *k-t* SENSE and compressed sensing theories.

k-t FOCUSS mainly aims to reduce the *(**k**, t)*-space sampling requirement without sacrificing the quality of the *(**x**, f)*-support [57]. Suppose $\boldsymbol{d}$ denotes the sampled data in the *(**k**, t)*-space and $\boldsymbol{\rho}$ denotes its corresponding support in the *(**x**, f)*-space. $\boldsymbol{d}$ is related to $\boldsymbol{\rho}$ by the spatial spectral Fourier transform $\boldsymbol{A}$,

$$\boldsymbol{d} = \boldsymbol{A}\boldsymbol{\rho}. \tag{3.26}$$

As is previously discussed, if a *(**k**, t)*-signal has sparse support in the *(**x**, f)*-space, un-aliased reconstruction of the original signal from the sparsely acquired *(**k**, t)*-samples can be realized by ℓ_1 minimization, even though the data acquisition speed is significantly below Nyquist sampling rate [68]. Given this theoretical guarantee in the compressed sensing theory, *k-t* FOCUSS aims to find a sparse solution defined as follows [57],

$$\begin{aligned} &\text{minimize } \|\boldsymbol{\rho}\|_1 \\ &\text{subject to } \|\boldsymbol{d} - \boldsymbol{A}\boldsymbol{\rho}\|_2 < \epsilon \end{aligned}\ , \tag{3.27}$$

where ϵ denotes noise level. *k-t* FOCUSS further expresses $\boldsymbol{\rho}$ as the product of $\boldsymbol{q}$ with a weighting matrix $\boldsymbol{W}$ that is renovated over multiple iterations [57],

$$\boldsymbol{\rho} = \boldsymbol{W}\boldsymbol{q}. \tag{3.28}$$

With equation (3.28), the original optimization problem can be transformed into the following problem [57],

$$\arg\min_{\boldsymbol{W}\boldsymbol{q}} \|\boldsymbol{d} - \boldsymbol{A}\boldsymbol{W}\boldsymbol{q}\|_2^2 + \lambda\|\boldsymbol{q}\|_2^2, \tag{3.29}$$

an optimal solution of which is given by [57],

$$\boldsymbol{\rho} = \hat{\boldsymbol{\rho}} + \boldsymbol{W}_n\boldsymbol{W}_n^H\boldsymbol{A}^H(\boldsymbol{A}\boldsymbol{W}\boldsymbol{W}^H\boldsymbol{A}^H + \lambda\boldsymbol{I})^{-1}(\boldsymbol{d} - \boldsymbol{A}\hat{\boldsymbol{\rho}}), \tag{3.30}$$

Where $\hat{\boldsymbol{\rho}}$ is the initial estimate of the ***(x**, f)*-image and $\boldsymbol{W}_n$ is the weighting matrix renovated in the n^{th} iteration [57],

$$\boldsymbol{W}_n = \begin{pmatrix} |\rho_{n-1}(1)|^p & 0 & \cdots & 0 \\ 0 & |\rho_{n-1}(2)|^p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & |\rho_{n-1}(N)|^p \end{pmatrix}, 0.5 \le p \le 1, \tag{3.31}$$

where the diagonal entries of $\boldsymbol{W}_n$ consists of $\boldsymbol{\rho}$ in the $(n\text{-}1)^{th}$ iteration and this weighting matrix allows for successive renovate of the sparse support in the ***(x**, f)*-space. Previous research has also indicated that when p = 0.5, the insignificant spectral coefficients will converge to zero while the significant ones are preserved, which suggests that the solution obtained by *k-t* FOCUSS can be regarded as an equivalent solution obtained through ℓ_1 minimization [74].

3.2 Partial separability model

3.2.1 Partially separable functions

The partially separable functions (PSFs) are a special class of functions that possess useful properties for dynamic MR imaging [8, 75]. As its name suggests, the definition of the partially separable function is based on the definition of complete separable functions. By its definition, a multivariate complete separable function can be decomposed as the product of simple functions of each individual variable [8, 75]. In dynamic imaging, for instance, if a speech signal is modeled as a separable function in space, its spatial variation can be decomposed as independent variation of each individual spatial variable. Without loss of generality, let us assume that the speech signal is modeled as a function of three variables, x_1, x_2, x_3,

$$\rho(x_1, x_2, x_3) = p_1(x_1)p_2(x_2)p_3(x_3). \tag{3.32}$$

Unlike complete separable functions, the PSF does not completely decompose a multivariate function into the product of multiple one-dimensional functions [8, 75]. Instead, it allows some variables to be partially combined,

$$\rho(x_1, x_2, x_3) = p(x_1, x_2)q(x_3). \tag{3.33}$$

The significance of the PSF lies in its potential to reduce the sampling requirement for dynamic speech imaging [8, 75]. According to the Nyquist theorem, a support limited signal in the ***(k**, t)*-space needs to be sampled at the Nyquist rate in order to be perfectly reconstructed. In practice, unfortunately, it is difficult to reach the Nyquist rate with a limited MR imaging speed since the number of Nyquist samples rises exponentially as the physical dimension of the problem increases. This difficulty is usually referred to as the *curse of dimensionality*, which usually compromises the spatiotemporal resolution of the reconstructed image [8, 75]. However, high spatiotemporal resolution can be realized with a sampling speed under the Nyquist rate when the de-

sired high-dimensional signal is modeled as a PSF, the required sampling rate of this PSF is usually much lower than that of the original high-dimensional signal since the PSF can effectively decrease the number of degrees of freedom [8, 75]. This favorable property of the PSF holds great potential in enabling sub-Nyquist sampling of the dynamic speech imaging signal.

The conventional definition of PSF has been further generalized to the L^{th}-order PSF [8, 75]. Compared with the conventional definition of PSF, the L^{th}-order PSF expresses a multivariate function as the sum of L basic PSFs. Mathematically, this can be expressed as,

$$\rho(x_1, x_2, x_3) = \sum_{l=1}^{L} p_{l,1}(x_1, x_2) p_{l,2}(x_3). \tag{3.34}$$

Compared with its primitive form, i.e. PSF, the L^{th} order PSF encompasses a broader class of signals by combining L groups of basic PSFs. A larger model order L usually suggests more complex signal dynamics can be captured and *vice versa*. This strong representation power of the L^{th} order PSF has been successfully exploited by the PS model to represent complex signal variations in the ***(k**, t)*-space [8, 75].

3.2.2 PS model induced low rank approximation

The PS model defines the ***(k**, t)*-space speech imaging signal on $K \times T$, where K denotes the spatial subspace, T denotes the temporal subspace and $K \times T$ denotes the Cartesian product of K and T [8, 75]. Using this model, the multidimensional oropharyngeal variations can be decomposed as partially separable temporal variations and spatial variations to the L^{th}-order. Mathematically, the PS model describes this relation with the following equation,

$$d(\boldsymbol{k}, t) = \sum_{l=1}^{L} c_l(\boldsymbol{k}) \varphi_l(t), \tag{3.35}$$

where $d(\boldsymbol{k}, t)$ denotes the measured data in the ***(k**, t)*-space, $\{c_l(\boldsymbol{k})\}_{l=1}^{L}$ denotes the spatial basis functions that represent the spatial variation and $\{\varphi_l(t)\}_{l=1}^{L}$ denotes the temporal basis functions that represent the temporal variations [8, 75]. By modeling the speech signal as L^{th} order PSF, a broad range of speech dynamics can be represented by changing the value of model order

L, which represents the level of spatiotemporal correlation of the desired image function [8, 75]. In other words, if a similar temporal pattern is shared among the majority of spatial locations, a small *L* is enough to describe their dynamics [8, 75]. On the contrary, if temporal characteristics vary among the majority of spatial locations, a large *L* is required. In other words, the representation power of the PS model can be deliberately controlled by selecting an appropriate model order [8, 75].

After discretization, the PS model is equivalent to a low-rank model of the speech dynamics [8, 75, 76]. This low-rank model is a good approximation of the dynamic speech imaging signals because it coincides with several important characteristics of speech imaging. Firstly, a specific speech motion is usually characterized by a limited number of sequential imaging frames. This suggests the fact that the speech articulators only need to sequentially visit a small number of vocal tract locations in order to complete a speech motion. As long as the sampled data can represent the articulator motion in these locations, no more information is needed to describe the entire speech production motion. Secondly, a speech image sequence only contains a limited number of moving pixels to describe articulator motion. For instance, in a typical mid-sagittal image of the upper vocal tract, a large number of image pixels describe static regions such as the brain and the cervical vertebra. Information on these static regions is actually redundant for the description of speech motion. Thirdly, the rank of speech dynamics is constrained by the number of driving muscles in the vocal tract. To produce a phonetically-meaningful sound, only a small number of muscles are needed to actuate articulator motion. In addition, these speech-driving muscles move in a highly correlated manner due to the fact that they are connected with joints and attachment points. In this way, a speech motion can be well characterized as long as the dominating speech motion is determined. These three characteristics of speech have made the PS model a good fit to reduce sampling requirements for dynamic speech imaging while maintaining fine details of articulator motion.

Although one may further argue that subtle speech dynamics are neglected by assuming a low-rank structure, the PS model offers great flexibility to adjust the level of desired speech dy-

namics. Since the spatiotemporal speech motion is expressed in the form of $\{d(\boldsymbol{k}_l, t)\}_{l=1}^{L}$, a larger model order L may help represent detailed speech dynamics. Previous research has also proven the representation power of the PS model with a variety of dynamic imaging applications, including cardiac MRI [77, 78] and MR spectroscopy [79, 80]. With these notable applications, the PS model holds great promise for capturing the major spatiotemporal events of normal speech production.

The PS model assumes that the desired speech image sequence exists in a $Q \times P$ dimensional space, where Q denotes the number of phase encodings, P denotes the number of imaging time frames [8, 75]. With this image model, we can rearrange the acquired speech data into a Casorati matrix in the following form [8, 75],

$$\boldsymbol{C} = \begin{bmatrix} d(\boldsymbol{k}_1, t_1) & d(\boldsymbol{k}_2, t_1) & \cdots & d(\boldsymbol{k}_Q, t_1) \\ d(\boldsymbol{k}_1, t_2) & d(\boldsymbol{k}_2, t_2) & \cdots & d(\boldsymbol{k}_Q, t_2) \\ \vdots & \vdots & \ddots & \vdots \\ d(\boldsymbol{k}_1, t_P) & d(\boldsymbol{k}_2, t_P) & \cdots & d(\boldsymbol{k}_Q, t_P) \end{bmatrix}, \tag{3.36}$$

where $\{\boldsymbol{k}_q\}_{q=1}^{Q}$ denotes sampling locations in the $\boldsymbol{k}$-space, $\{t_p\}_{p=1}^{P}$ denotes the sampling time point. The rank of the Casorati matrix is assumed to be L. Previous research has demonstrated that the Casorati matrix $\boldsymbol{C}$ has a maximum of $2(P + Q - L)L$ degrees of freedom, which is significantly smaller than the number of entries in $\boldsymbol{C}$ [81]. With small degrees of freedom, sub-Nyquist rate sampling can be utilized to accelerate the imaging speed.

The PS model attempts to extract the temporal basis functions $\{\varphi_l(t)\}_{l=1}^{L}$ from a data set that satisfies temporal Nyquist rate. Usually, this data set is referred to as the navigator data set. Specifically, the acquired navigator data are rearranged into a Casorati matrix $\boldsymbol{C}$ in equation (3.36) and the temporal basis functions $\{\varphi_l(t)\}_{l=1}^{L}$ can be determined directly from the column space of $\boldsymbol{C}$ [8, 75]. In this thesis, the singular value decomposition (SVD) is applied to find the temporal basis functions $\{\varphi_l(t)\}_{l=1}^{L}$. Mathematically, this can be expressed as

$$\boldsymbol{C} = \sum_{l=1}^{\min\{P,Q\}} \boldsymbol{\mu}_l \sigma_l \boldsymbol{v}_l^{\mathrm{H}}, \tag{3.37}$$

where *H* denotes the Hermitian transpose, $\{\sigma_l\}_{l=1}^{L}$ denotes the singular values arranged in descending order, $\{\boldsymbol{\mu}_l\}_{l=1}^{L}$ denotes the left singular vectors and $\{\boldsymbol{v}_l\}_{l=1}^{L}$ denotes the right singular vectors. Usually the *L*-most-significant left singular vectors from SVD are taken as the temporal basis functions [8, 75],

$$\boldsymbol{\mu}_l = \{\varphi_l(t_1), \varphi_l(t_2), \dots, \varphi_l(t_P)\}^{\mathrm{T}}. \tag{3.38}$$

It is worth noting that there exist other methods to extract the temporal basis function, including the usage of complex exponentials [82], as well as the use of a low-order ARMA model [83]. In this thesis, however, we focus on extracting the temporal basis function from the navigator data set using SVD given its notable applications in previous studies [84, 85].

3.3 PS Model-based data acquisition

3.3.1 Common characteristics of PS model-based sampling scheme

The PS model adopts a composite data acquisition scheme in the ***(k**, t)*-space. Two data sets are acquired in order to estimate the temporal basis functions and the spatial basis functions [8, 75]. These two data sets are the navigator data set that has high temporal resolution, as well as the imaging data set that has high resolution spatial information [8, 75]. Since the PS model assumes that the temporal subspace and the spatial subspace are partially separable, data acquisition of the navigator data set is not constrained by the spatial Nyquist criterion [8, 75]. The navigator data set is effectively acquired as long as the temporal Nyquist rate is satisfied. Similarly, data acquisition of the imaging data set is not limited by the temporal Nyquist criterion [8, 75]. The imaging data set is effectively acquired in a way that the ensemble of sampled ***k***-space locations satisfies the spatial Nyquist rate [8, 75]. With the above composite data acquisition scheme, the sampling requirement for high-dimensional ***(k**, t)*-space data can be effectively reduced.

In practice, there exist some specific requirements on how the navigator data set and the imaging data set are acquired. These requirements are mainly based on algorithm and implementation considerations. Take the algorithm consideration as an example. Each $\boldsymbol{k}$-space location should be sampled at least L times in the imaging data set to avoid solving an underdetermined problem [75]. Unlike the algorithm considerations, the implementation considerations attempt to reduce RF energy deposition to the speech subject, minimize stimulation to the human peripheral nerve system and improve eddy current performance of the gradient system [75]. Other considerations may depend on the specific imaging application [8, 75]. Despite these detailed requirements, the PS model sampling scheme in general allows for a high level of freedom in designing acquisition patterns for both the navigator data set and the imaging data set. Figure 3.7 illustrates a time-sequential acquisition of these two data sets in the $(\boldsymbol{k}, t)$-space.

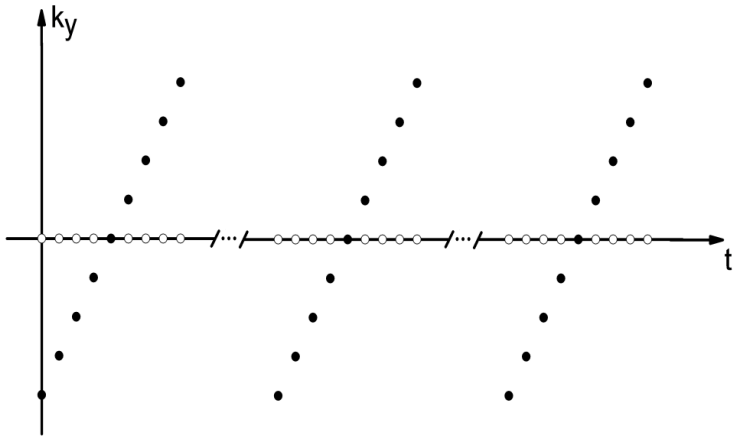


Figure 3.7: Illustration of the time-sequential composite acquisition scheme for the PS model. Each dot in the above figure represents the acquisition of one phase encoding line in the $\boldsymbol{k}$-space, where the frequency encoding axis is omitted for simplicity. White dots indicate the acquisition of the navigator data set, which is acquired with high temporal resolution. Black points indicate the acquisition of the imaging data set, which is acquired with high spatial resolution.

3.3.2 Dependency of temporal dynamics on navigator placement

Temporal dynamics of the PS model-based reconstruction is closely related to the navigator data

set. The navigator data set determines the PS model temporal subspace, which uniquely characterizes the temporal variations of the dynamic signal. Also, as will be discussed later in Section 3.4.1, the spatial basis functions $\{c_l(\boldsymbol{k})\}_{l=1}^{L}$ are determined from least-square inversion in which both the imaging data set and the temporal basis functions are involved [8, 75]. In this way, an incorrect estimation of the temporal subspace not only hampers temporal resolution, but also degrades spatial resolution in the reconstructed image sequences. Therefore, it is important to obtain accurate estimation of temporal subspace by carefully designing the navigator data set.

The influence of the navigator data on the temporal subspace is governed by two factors: the placement and the orientation of the navigator sampling pattern [8, 75]. The navigator placement refers to the locations of the sampled data points in the ***(k**, t)*-space. Ideally, if the ***(k**, t)*-space is fully covered by the navigator, all temporal basis functions that span the temporal subspace can be extracted from the collected data. In reality, however, some temporal basis functions may be "skipped" since the ***(k**, t)*-space is sparsely sampled. Insufficient temporal basis functions prevent the temporal dynamics to be faithfully reconstructed [8, 75]. Therefore, the navigator placement should ideally cover the ***(k**, t)*-locations where the non-negligible energy components reside. On the other hand, the navigator orientation mainly refers to the angles covered by the navigator trajectories. Conventionally the Cartesian navigators are used to sweep through the ***k***-space center at every imaging time frame. However, Cartesian lines could fail to capture motion in the orthogonal direction [75]. In order to capture articulator motion, the navigator trajectory should be designed to distribute data samples among multiple orientations. This is how the navigator placement and navigator trajectory influences the estimation of temporal subspace.

3.3.3 Design of alternative navigator sampling schemes

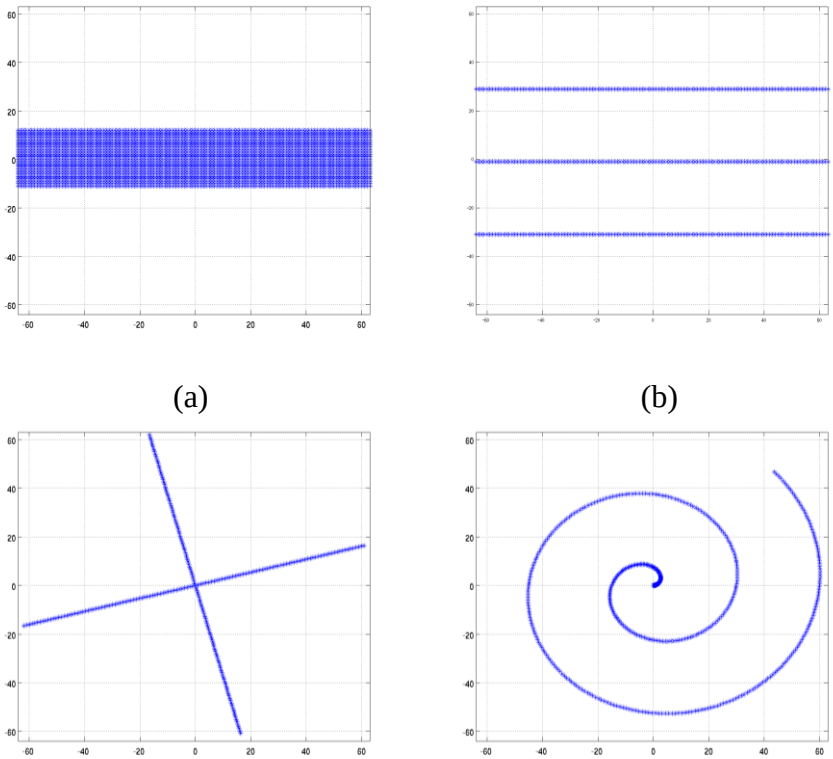
To optimize the navigator sampling location and sampling orientation, two categories of navigator sampling patterns are employed to examine their influence on reconstruction. The first category of navigator sampling patterns aims to determine whether increasing navigator sampling

locations would lead to better spatiotemporal dynamics. To simplify our investigation, the MR imaging speed is not taken into account and a series of ideal Cartesian trajectories are used for investigation. The second category of navigator sampling patterns aims to determine whether broader orientation coverage would lead to better speech dynamics. Since the Cartesian trajectories would fail to capture speech dynamics in its orthogonal direction [75], the non-Cartesian trajectories are introduced in this category. Non-Cartesian trajectories include the radial trajectories and the spiral trajectories. In addition, the design of the trajectories in the second category is subject to the typical time constraint in dynamic imaging experiments. As previously discussed in Chapter 2, state-of-the-art speech analysis requires a temporal resolution of 20 fps or more across multiple imaging planes. According to our previous investigations on speech imaging [86], we assume a typical 10 ms TR for all the navigator trajectories to perform a fair comparison. Under this time constraint, three Cartesian lines are used for Cartesian trajectories, three projection lines are used for radial trajectories, and two spiral turns are used for spiral trajectories.

Overall this thesis investigates the performance of four categories of navigator sampling patterns. Within each category, the navigator trajectories are modified in an attempt to capture the slight differences in spatiotemporal dynamics. Specifically, Fig. 3.8(a) depicts an ideal Cartesian trajectory that has the maximum spatial coverage. In this category, the navigator sampling pattern varies between 2, 4, 8, 16 and 24 Cartesian lines. Without loss of generality, these Cartesian lines are all placed around the ***k***-space center. The word "ideal" refers to the fact that the acquisition of most trajectories (24 Cartesian lines, for instance) exceeds the 10 ms time constraint. The purpose of this category of navigator sampling patterns is to observe the change in the reconstructed articulator dynamics with regard to ***k***-space coverage.

Figure 3.8(b) depicts a conventional Cartesian trajectory used for comparison. The word "conventional" refers to the fact that all the Cartesian trajectories in this category can be acquired within 10 ms. These Cartesian trajectories are placed in five distinctive patterns. These patterns are designed to cover the ***k***-space in a way that the center, the high-frequency region, the mid-frequency region, the center and the high-frequency region, the center and the mid-frequency

regions are detected. The purpose of placing Cartesian navigators in these regions is to determine whether the high-frequency regions in ***k***-space may facilitate preserving the fast varying localized speech dynamics. Figure 3.8(c) depicts a radial trajectory used for comparison. In this category, the radial trajectories cover the projection angles from 0 °, 15 °, 30 °, 60 ° and 75 °. The design of these trajectories attempts to divide the ***k***-space into regions covered by 15 ° projection angle. The purpose of these navigator placements is to demonstrate the effect of navigator orientation on the reconstruction. Figure 3.8(d) depicts a spiral trajectory used for comparison. Similarly with the radial trajectories, these spiral trajectories aim to demonstrate the effect of navigator orientation on the reconstruction. Specifically, these spiral trajectories cover rotation angles 0 °, 45 °, 90 °, 135 ° and 180 °. Unlike radial trajectories, the ***k***-space sampling density of spiral trajectories can be arbitrarily adjusted. Therefore, three additional spiral trajectories are added for comparison with varied density in the ***k***-space center and the ***k***-space edge.



(a) (b)

(c) (d)

Figure 3.8: An example of the (a) ideal Cartesian sampling pattern (acquisition exceeds 10 ms), (b) Cartesian sampling pattern, (c) radial sampling pattern and (d) spiral sampling pattern.

3.4 PS model-based image reconstruction

3.4.1 Basic PS reconstruction

Image reconstruction of dynamic image sequences assumes that the temporal basis functions are already obtained accurately from the navigator data set [8]. Therefore, the main goal of basic PS reconstruction is to extract the spatial basis functions from an imaging data set that has high resolution spatial information [8]. Mathematically, extraction of the spatial basis functions is usually modeled as a least-square problem [8],

$$\left\{\widehat{c_l(\boldsymbol{k})}\right\}_{l=1}^{L} = \underset{\{c_l(\boldsymbol{k})\}_{l=1}^{L}}{\arg\min} \|d(\boldsymbol{k},t) - \sum_{l=1}^{L} c_l(\boldsymbol{k})\varphi_l(t)\|^2, \tag{3.39}$$

where $d(\boldsymbol{k},t)$ denotes the imaging data set, $c_l(\boldsymbol{k})$ denotes the spatial basis functions and $\varphi_l(t)$ denotes the temporal basis functions. Solving the optimization problem in equation (3.39) is equivalent to obtaining the spatial basis functions from the following equation,

$$\begin{bmatrix} \varphi_1(t_{1,q}) & \varphi_2(t_{1,q}) & \cdots & \varphi_L(t_{1,q}) \\ \varphi_1(t_{2,q}) & \varphi_2(t_{2,q}) & \cdots & \varphi_L(t_{2,q}) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(t_{P,q}) & \varphi_2(t_{P,q}) & \cdots & \varphi_L(t_{P,q}) \end{bmatrix} \begin{bmatrix} c_1(\boldsymbol{k}_q) \\ c_2(\boldsymbol{k}_q) \\ \vdots \\ c_L(\boldsymbol{k}_q) \end{bmatrix} = \begin{bmatrix} d(\boldsymbol{k}_q, t_{1,q}) \\ d(\boldsymbol{k}_q, t_{2,q}) \\ \vdots \\ d(\boldsymbol{k}_q, t_{P,q}) \end{bmatrix}, \tag{3.40}$$

where q = 1, 2, …, Q indexes each $\boldsymbol{k}$-space location $\boldsymbol{k}_q$, p = 1, 2, …, P indexes each imaging time frame t_p. To simplify expression, the above matrix equation can usually be written in the following abstract form [8],

$$\boldsymbol{\Phi}_q \boldsymbol{\Psi}_q = \boldsymbol{d}_q. \tag{3.41}$$

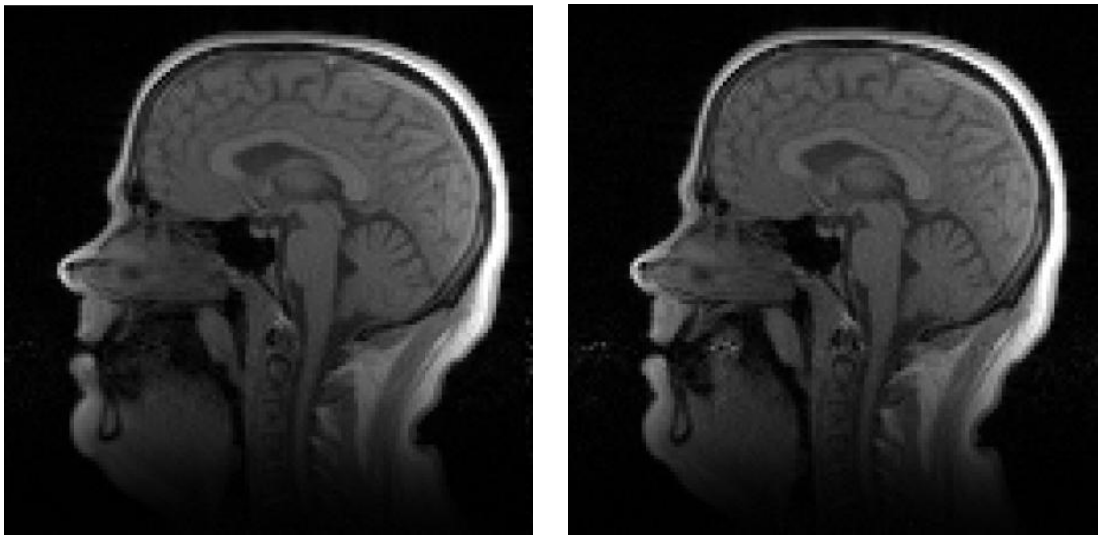
In equation (3.41), if the number of imaging time frames P is larger than the model order L, the equation is obviously a well-determined system, the solution of which can be obtained from least-square inversion [8]. When both the temporal basis functions and the spatial basis functions are known, image reconstruction is usually completed by summing L products of $c_l(\boldsymbol{k})$ and $\varphi_l(t)$

according to the model formulation [8].

As discussed in the previous sections, the ability of the reconstructed image to represent oropharyngeal dynamics depends largely on the selection of model order L [8]. If the value of L is too small, an insufficient number of spatial and temporal basis functions are chosen to describe spatiotemporal variations, even though the matrix inversion problem has good conditioning [8, 75]. A limited number of spatial and temporal basis functions reduces the level of localized signal dynamics in the reconstructed images and hence introduces blurring in both the spatial and temporal domains [75]. This can be exemplified in Fig 3.9(a) where a model order of 5 is used to capture speech dynamics. In this figure, the blurred motion of the tongue indicates that the number of temporal and spatial basis functions is insufficient to represent complex speech motion.

However, if L is assigned a large value and the imaging data set is highly undersampled, the least-square fitting problem is ill-conditioned [84, 87]. In other words, the reconstructed image may be compromised by ill-conditioning issues, such as greater noise or fraudulent motion pattern, even though the underlying least-square inversion problem is well-posed [84, 87]. This can be exemplified by Fig. 3.9 (b) where a model order of 25 is used to capture speech dynamics. Although this figure demonstrates contact of the tongue tip with the hard palate, speech dynamics in the oral cavity is largely contaminated by the amplified noise. The amplified noise compromises the shaping of the tongue and the velum in the oral cavity and the velum cavity.

Although high model order is desired for capturing detailed oropharyngeal dynamics in dynamic speech imaging, the resultant ill-conditioning effects and image artifacts prevent the application of least-square inversion to extract the spatial basis functions [84, 87]. In order to suppress ill-conditioning, previous research in our group have applied spatial spectral support constraints to regularize the model fitting problem [84, 87]. Given these notable applications, this spatial spectral constraint is applied in combination with the basic PS reconstruction to assist reconstruction.



(a) (b)

Figure 3.9: (a) Basic PS reconstruction using a small model order (L = 5). (b) Basic PS reconstruction using a large model order (L = 25).

3.4.2 Basic PS reconstruction with sparsity constraint

Previous research in [85] has proposed to regularize ill-conditioned data fitting associated with basic PS reconstruction through mutually imposing the partial separability constraint and the spatial spectral sparsity constraint. This method has already found notable applications in the context of cardiac imaging [85]. Therefore, it is natural to extend this method to other dynamic imaging applications that share similarity with cardiac imaging.

Dynamic speech imaging is similar to cardiac imaging in that the desired imaging signal is also sparse in the $(\boldsymbol{x}, f)$-space. Sparsity in the $(\boldsymbol{x}, f)$-space is reflected by two important characteristics of dynamic speech imaging signal. Firstly, only a small number of pixels in the dynamic image sequences describe articulator movements in space [85]. A larger number of pixels, on the contrary, describe the background that remains static in each imaging frame. In this way, the static portion of the dynamic image sequences can be compressed and sparsely represented [85]. Secondly, some desired speech motion is periodic or quasi-periodic. Periodicity of speech motion also leads to a higher level of sparsity in the temporal frequency spectrum [85]. With these two

characteristics, the dynamic speech image signal can be transformed into a series of coefficients, in which only a few non-zero elements exist [88]. In the transform domain, specifically, the number of non-zero elements is usually measured with the l_0 norm. Therefore, imposing sparsity on the dynamic speech images is equivalent to minimizing the l_0 norm in the ***(x, f)***-space.

Specifically, the spatial spectral sparsity constraint can be imposed on the PS model to suppress image artifacts induced by ill-conditioning [89]. Based on previous discussion, the sparsity constraint is imposed by minimizing the l_0 norm in the transform domain. However, direct optimization of the l_0 norm is unrealistic [90]. Fortunately, the l_1 norm can be applied to replace the original l_0 norm and transform the l_0 norm minimization problem to an alternative problem that can be settled via existing mathematical methods [91]. In this thesis, specifically, the l_1 norm is used to impose ***(x, f)***-sparsity on the dynamic speech imaging signal. Correspondingly, the sparsifying transform is chosen to be the temporal Fourier transform. In this way, the optimization problem can be written as,

$$\widehat{\mathbf{C}_s} = \underset{\mathbf{C}_s \in \mathbb{C}^{\mathrm{Q}\times\mathrm{L}}}{\arg\min} \|\mathbf{d} - \Phi\{\mathbf{F}_s \mathbf{C}_s \boldsymbol{\Psi}_t\}\|_2^2 + \lambda \|\mathbf{C}_s \boldsymbol{\Psi}_t \mathbf{F}_t\|_1 , \tag{3.42}$$

where Φ: $\mathbb{C}^{Q\times P} \rightarrow \mathbb{C}^{M\times 1}$ denotes a sparse sampling operator in the ***(k, t)***-space, $\mathbf{d} \in \mathbb{C}^{M\times 1}$ denotes the measured data, $\mathbf{C}_s \in \mathbb{C}^{Q\times L}$ denotes a basis for the spatial subspace, $\boldsymbol{\Psi}_t \in \mathbb{C}^{L\times P}$ denotes the matrix that holds a basis for the temporal subspace, $\mathbf{F}_s \in \mathbb{C}^{Q\times Q}$ denotes a spatial Fourier transform matrix, $\boldsymbol{F}_t \in \mathbb{C}^{P\times P}$ denotes a temporal Fourier transform matrix and λ denotes the regularization parameter. The above formulation has been previously developed in [92] to encompass both the partial separability constraint and the spatial-spectral sparsity constraint into one composite expression.

The above composite expression provides a great deal of freedom in choosing the desired constraint for reconstruction [92]. For instance, when $\lambda = 0$, this formulation can be considered as a basic PS reconstruction problem [92]. When L = M, however, this formulation can be regarded as sparsity-constrained reconstruction [92]. Basic sparse reconstruction utilizes ***(x, f)***-sparsity of the speech imaging signal but previous research in our group indicates that this

method often experiences motion blurring problems when the ***(k**, t)*-space is highly undersampled [92]. With an appropriate value of L and λ, the partial separability constraint and spatial-spectral sparsity constraint can be mutually imposed in the hope that reconstruction quality can be refined from the complementary interaction between both constraints [92].

An algorithm based on half-quadratic regularization with continuation processes has been previously proposed to efficiently settle the above optimization problem [85]. Since the l_1 norm is not differentiable at zero, the l_1 norm in the above optimization problem is approximated by the Huber function [85]. Mathematically, this can be expressed as,

$$\widehat{\mathbf{C}_s} = \underset{\mathbf{C}_s \in \mathbb{C}^{\mathrm{Q}\times\mathrm{L}}}{\arg\min} \|\mathbf{d} - \Phi\{\mathbf{F}_s \mathbf{C}_s \boldsymbol{\Psi}_t\}\|_2^2 + \lambda \sum_{q=1}^{Q} \sum_{p=1}^{P} \varphi\left\{\left|(\mathbf{C}_s \boldsymbol{\Psi}_t \mathbf{F}_t)_{q,p}\right|\right\}, \tag{3.43}$$

where the Huber function $\varphi(h)$ is introduced to approximate $|h|$. The Huber function is parameterized by the scalar α,

$$\varphi(h) = \begin{cases} \frac{|h|^2}{2\alpha}, & \text{if } |h| \le \alpha, \\ |h| - \frac{\alpha}{2}, & \text{if } |h| > \alpha. \end{cases} \tag{3.44}$$

The accuracy of Huber function approximation is determined by the value of the scalar α. As α approaches zero, the Huber function closely approximates the non-differentiable l_1 norm [85, 92]. The Huber function can be understood from another perspective. It can be regarded as a combination of the l_1 norm and the l_2 norm [85, 92],

$$\varphi(h) = \min_g \left\{\frac{(h-g)^2}{2\alpha} + |g|\right\}, \tag{3.45}$$

where g denotes an supplementary variable on $\mathbb{R}$. With the introduction of Huber function in equation (3.45), the original optimization problem can be rewritten as [85, 92],

$$\left\{\widehat{\mathbf{C}_s}, \widehat{\mathbf{G}}\right\} = \underset{\mathbf{C}_s \in \mathbb{C}^{\mathrm{Q}\times\mathrm{L}}, \mathbf{G} \in \mathbb{C}^{\mathrm{Q}\times\mathrm{P}}}{\arg\min} \|\mathbf{d} - \Phi\{\mathbf{F}_s \mathbf{C}_s \boldsymbol{\Psi}_t\}\|_2^2 + \frac{\lambda}{2\alpha} \|\mathbf{C}_s \boldsymbol{\Psi}_t \mathbf{F}_t - \mathbf{G}\|_F^2 + \lambda \|\mathbf{G}\|_1. \tag{3.46}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm and **G** denotes an supplementary matrix. Equation (3.46) suggests that, with a fixed value of α, the original optimization problem can be transformed into an equivalent optimization problem [85, 92]. Previous research in our group has developed an

efficient algorithm to settle this equivalent problem based on half-quadratic regularization with continuation processes [85, 92].

Specifically, the supplementary matrix $\mathbf{G}$ and the spatial basis matrix $\mathbf{C}_s$ are alternatively optimized in each iteration by minimizing one with the other fixed [85, 92]. Assume that $\mathbf{C}_{\mathrm{S}}^{(w-1)}$ is already fixed in the current w^{th} iteration, $\mathbf{G}^{(w)}$ can thus be determined in the following way [85, 92],

$$\mathbf{G}_{i,j}^{(w)} = \begin{cases} 0 & , \text{ if } \left|\mathbf{C}_{\mathrm{S}}^{(w-1)}\boldsymbol{\Psi}_f\right|_{i,j} \le \alpha, \\ \frac{(\mathbf{C}_{\mathrm{S}}^{(w-1)}\boldsymbol{\Psi}_f)_{i,j}}{\left|\mathbf{C}_{\mathrm{S}}^{(w-1)}\boldsymbol{\Psi}_f\right|_{i,j}}\left(\left|\mathbf{C}_{\mathrm{S}}^{(w-1)}\boldsymbol{\Psi}_f\right|_{i,j} - \alpha\right), & \text{if } \left|\mathbf{C}_{\mathrm{S}}^{(w-1)}\boldsymbol{\Psi}_f\right|_{i,j} > \alpha, \end{cases} \tag{3.47}$$

where $\mathbf{C}_{\mathrm{S}}^{(w-1)}\boldsymbol{\Psi}_f \in \mathbb{C}^{P\times Q}$, $\mathbf{G}^{(w)}$ is renovated through an element-wise operation. With an renovated $\mathbf{G}^{(w)}$, $\mathbf{C}_{\mathrm{S}}^{(w)}$ can be in turn updated by minimizing the following formulation [85, 92],

$$\widehat{\mathbf{C}_s^{(w)}} = \underset{\mathbf{C}_{\mathrm{S}}^{(w)}\in\mathbb{C}^{\mathrm{Q}\times\mathrm{L}}}{\arg\min} \|\mathbf{d} - \Phi\{\mathbf{F}_s\mathbf{C}_s\boldsymbol{\Psi}_t\}\|_2^2 + \frac{\lambda}{2\alpha}\left\|\mathbf{C}_s\boldsymbol{\Psi}_f - \mathbf{G}^{(w)}\right\|_F^2. \tag{3.48}$$

$\mathbf{G}$ and $\mathbf{C}_s$ are alternatively renovated in each iteration until they reach convergence, the speed of which is mainly characterized by the Huber function scalar α. Overall, the value of α gradually decreases as the number of iterations increases [92]. In early iterations, α is usually assigned a relatively large value to guarantee fast convergence [92]. In the iterations that follow, the value of α is gradually reduced to yield a good approximation of the l_1 norm. Previous research in our group has guaranteed convergence of this algorithm [92]. In addition, a special structure has also been developed to accelerate computation on a row-by-row basis [85].

CHAPTER 4

RESULTS AND DISCUSSION

In this chapter, an implementation of basic PS reconstruction is applied to the dynamic speech problem to investigate the performance of this method in capturing articulator movements. In addition, a numerical phantom is generated to investigate the dependency of the performance of reconstructions on the navigator sampling pattern.

4.1 Simulations

4.1.1 Numerical phantom for dynamic speech imaging

A two-dimensional complex-valued numerical phantom was developed for dynamic speech imaging simulations. To build this numerical phantom, specifically, the underlying data were acquired from a real speech experiment that captures speech production of repetitive /za/-/na/ sounds from a Siemens Trio 3T scanner with a 12-channel receiver coil. The acquired data set covered a 280 mm × 280 mm × 40 mm FOV encompassing major articulators and the human brain in a mid-sagittal slice. Key frames representing important motion of the /za/-/na/ sounds were selected from the reconstruction of the acquired data. Non-rigid transformations were used in post-processing to interpolate the selected key frames into a high-temporal-resolution phantom data set. This numerical speech phantom has two properties: 1) the phantom data can mimic ap-

propriate articulator motion and 2) the phantom data have high spatiotemporal resolution.

Since a limited number of key frames were selected from the reconstruction of experimental data, interpolation was needed to create intermediate image frames for a high spatiotemporal resolution phantom. For a complex-valued phantom, interpolation is performed in a way that smooth variation is created across every intermediate frame. Specifically, temporal variation of articulator motion was created from the thin-plate spline (TPS) transform [93]. The TPS transform serves as an effective tool for temporal interpolation and it has been previously used to create smooth inter-frame variations for any physiological motion [94 - 97]. The TPS transform characterizes a physiological motion between two image frames with changes in the relative spatial location of two sets of control points [94 - 97]. These control points are usually referred to as landmark points, which can be placed around the speech articulators to represent speech motion [94 - 97]. Let us assume that control points (x_n^1, y_n^1) and (x_n^2, y_n^2) are determined on the image frame ρ_1 and ρ_2 that represent articulator motion, respectively. The purpose of TPS transform, therefore, is to determine two distinctive transforms $T_1(\cdot)$ and $T_2(\cdot)$ to match with the intermediate image frame ρ_3 between ρ_1 and ρ_2 [94 - 97].

In general, the TPS transform is carried out in two major steps [97]. The first step is to determine two distinctive spatial-transform functions $f(\cdot)$ and $g(\cdot)$ for ρ_1 and ρ_2 [94 - 97],

$$\begin{cases} x_n^1 = f(x_n^2, y_n^2) \\ y_n^1 = g(x_n^2, y_n^2) \end{cases}, \tag{4.1}$$

where $f(\cdot)$ and $g(\cdot)$ are the targeted spatial-transform functions that combine both affine transform and TPS transform in their expression [94 - 97],

$$\begin{cases} f(x,y) = a_1 + a_2 x + a_3 y + \sum_{n=1}^{N} c_n P\big((x,y),(x_n,y_n)\big) \\ g(x,y) = b_1 + b_2 x + b_3 y + \sum_{n=1}^{N} d_n P\big((x,y),(x_n,y_n)\big) \end{cases}, \tag{4.2}$$

where $P\big((x_m,y_m),(x_n,y_n)\big) = \|(x_m,y_m)-(x_n,y_n)\|^2 \cdot \log\|(x_m,y_m)-(x_n,y_n)\|^2$ is the TPS kernel function [94 - 96]. Previous research has found the relationship between the control points and the TPS kernel function [97],

$$\begin{bmatrix} \boldsymbol{P} & \boldsymbol{Q} \\ \boldsymbol{Q}^T & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{c} & \boldsymbol{d} \\ \boldsymbol{a} & \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x^1} & \boldsymbol{y^1} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \tag{4.3}$$

where

$$\boldsymbol{x^1} = (x_1^1, x_2^1, \cdots, x_N^1)^T,$$

$$\boldsymbol{y^1} = (y_1^1, y_2^1, \cdots, y_N^1)^T,$$

$$\boldsymbol{x^2} = (x_1^2, x_2^2, \cdots, x_N^2)^T,$$

$$\boldsymbol{y^2} = (y_1^2, y_2^2, \cdots, y_N^2)^T,$$

$$\boldsymbol{Q} = (\boldsymbol{1}, \boldsymbol{x^2}, \boldsymbol{y^2}),$$

$$\boldsymbol{P}(m,n) = P\big((x_m^2, y_m^2), (x_n^2, y_n^2)\big).$$

With equation (4.3), the relation between ρ_1 and ρ_2 can be defined when the spatial-transform functions $f(\cdot)$ and $g(\cdot)$ are determined. The second step, however, is to determine ρ_3 from the spatial-transform functions $f(\cdot)$ and $g(\cdot)$ [94 - 97]. Assume that a set of control points (x_n^3, y_n^3) is defined on ρ_3, the relation between (x_n^3, y_n^3) and the previous two sets of control points, (x_n^1, y_n^1) and (x_n^2, y_n^2), can be expressed with the following equation [97],

$$\begin{cases} x_n^3 = (1-\lambda_t)x_n^1 + \lambda_t x_n^2 \\ y_n^3 = (1-\lambda_t)y_n^1 + \lambda_t y_n^2 \end{cases}, \tag{4.4}$$

where λ_t is a parameter defined between 0 and 1 that describes linear spatiotemporal variation With the above equation, the intermediate frame ρ_3 can be fully characterized when λ_t is determined [97]. Usually, λ_t is defined by assuming that speech dynamics changes linearly with time [97] and it can be determined by the following equation [97],

$$\lambda_t = \frac{t_2 - t_3}{t_2 - t_1}. \tag{4.5}$$

This linear relation has also been previously applied to successfully perform non-rigid registration and is also applied in this thesis to approximate natural articulator motion [97]. Figure 4.1 depicts the interpolated image frames between two reference frames via the TPS transform.

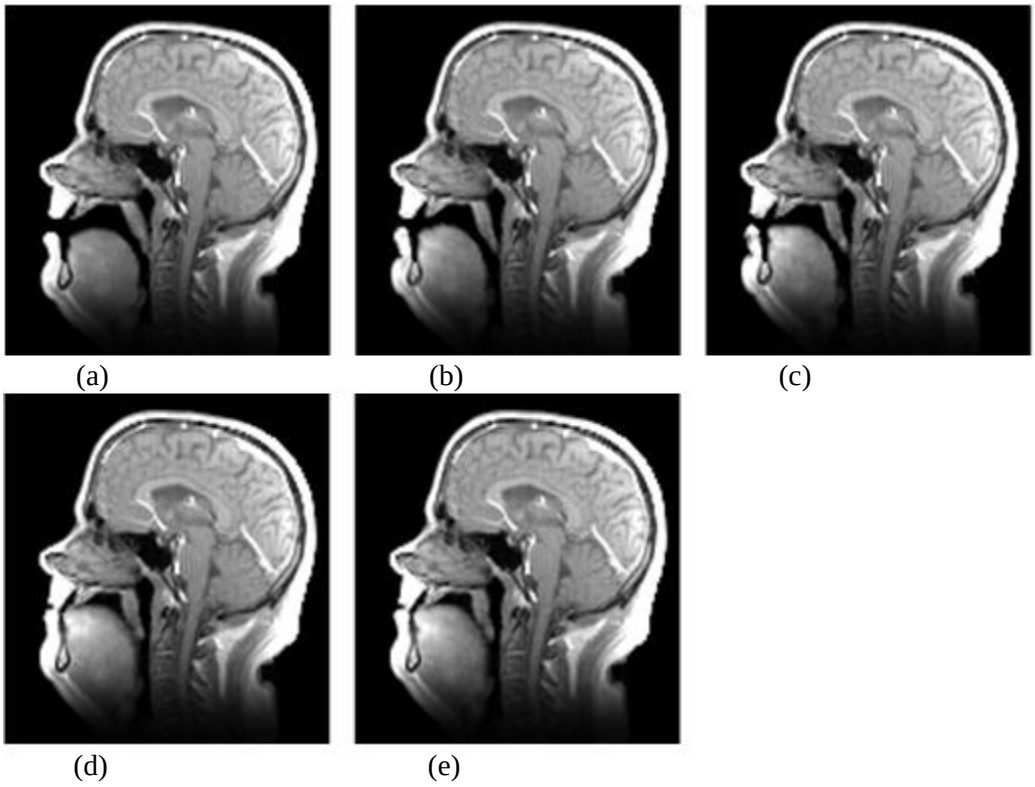


Figure 4.1: The reference image frames (a), (e) and the interpolated image frames (b), (c), (d).

4.1.2 Comparison in terms of navigator sampling patterns

4.1.2.1 Quantitative and qualitative metrics for comparison

The speech phantom generated from the TPS transform is used to simulate PS model-based data acquisition and image reconstruction schemes. For simplicity, the sampling pattern for the imaging data set is fixed as time-sequential Cartesian sampling, while the sampling patterns for the navigator data sets vary. The sampled data are reconstructed with the basic PS algorithm without regularization.

The analysis on the temporal subspace is carried out in three steps. Firstly, a gold standard dynamic image sequence is reconstructed from the fully sampled navigator data and the time-sequentially sampled imaging data set through the PS model. Although the fully sampled

navigator data is unrealistic in practice, the reconstructed gold standard image sequence provides a reference for further investigation. Secondly, dynamic image sequences are also reconstructed from the navigator data set sampled with alternative navigator patterns mentioned in Section 3.3.3. Thirdly, the difference between these image sequences and the gold standard is measured by a variety of quantitative and qualitative metrics. These metrics include the normalized root mean square error (NRMSE), the time average mean square error (TAMSE), the condition number, the error map and the strip plot. These comparisons are helpful for indicating which navigator pattern yields minimum loss in speech dynamics.

The NRMSE is used to quantitatively measure global and local differences between the gold standard image and the reconstructed image [81]. The NRMSE serves as an indicator of the overall reconstruction error and can be calculated in the following way [81],

$$\text{NRMSE} = \sqrt{\frac{\sum_{p=1}^{P}\sum_{q=1}^{Q}\left|\rho(\mathbf{r}_q,t_p)-\rho_g(\mathbf{r}_q,t_p)\right|^2}{\sum_{p=1}^{P}\sum_{q=1}^{Q}\left|\rho_g(\mathbf{r}_q,t_p)\right|^2}} \tag{4.6}$$

where $\boldsymbol{r}_q = (\mathrm{x}_q, \mathrm{y}_q)^{\mathrm{T}}$ denotes the spatial coordinate, t_p denotes the p^{th} imaging frame, $\rho_g(\mathbf{r}_q, t_p)$ denotes the value of $\boldsymbol{r}_q$ in the p^{th} imaging frame of the gold standard image sequence, $\rho(\mathbf{r}_q, t_p)$ denotes the value of $\boldsymbol{r}_q$ in the p^{th} imaging frame of the image sequence reconstructed from different navigator sampling patterns. The above formulation sums up the image differences of every pixel across all imaging frames and hence indicates quantitatively the overall reconstruction quality. In addition, the region in which the NRMSE is calculated can be arbitrarily determined. Since the dynamics of the speech articulators are limited to a small region of the image, in this thesis we confine the definition of NRMSE to a local region where articulator movements are most likely to take place. Figure 4.2 depicts this local region with a dotted rectangular box. As can be seen in Fig. 4.2, the dotted rectangular box includes major articulators in the vocal tract, such as the lip, the palate, the tongue, the velum, the epiglottis and the pharyngeal wall. The NRMSE value in this region mainly reflects the localized image difference and is sensitive to the reconstruction error caused by different navigator sampling patterns.

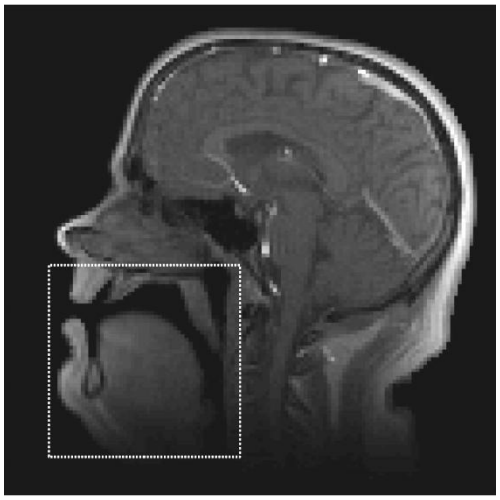
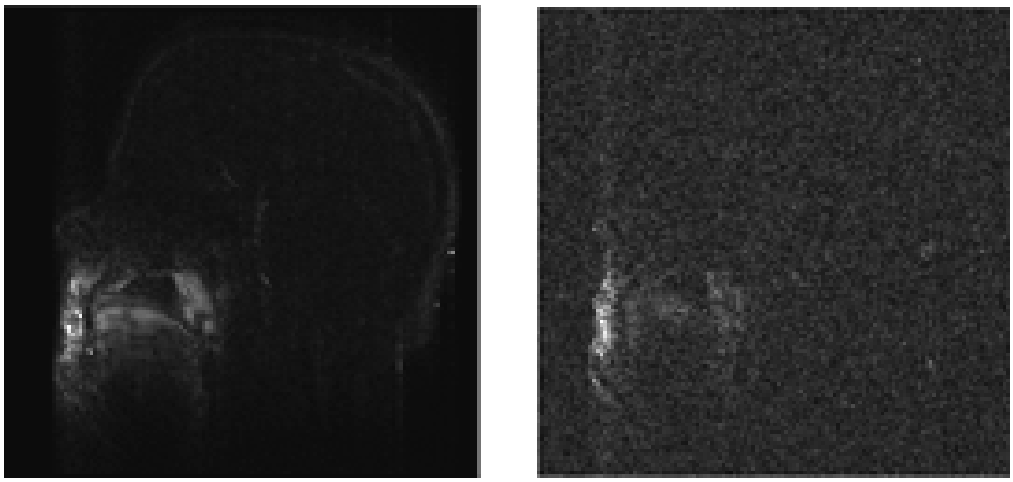


Figure 4.2: The local region that includes major vocal articulators.

Unlike the NRMSE, the TAMSE does not measure the overall image difference. Instead, the TAMSE presents a map that describes the spatially-dependent distribution of accumulated image difference over time. The TAMSE for an arbitrary pixel $\rho(\mathrm{x}_0, \mathrm{y}_0)$ in an image can be usually defined as,

$$\mathrm{TAMSE}_{\mathbf{r}_q} = \frac{1}{P}\sqrt{\frac{\sum_{p=1}^{P}|\rho(\mathbf{r}_q,t_p)-\rho_g(\mathbf{r}_q,t_p)|^2}{\sum_{p=1}^{P}|\rho_g(\mathbf{r}_q,t_p)|^2}}, \tag{4.7}$$

where $\rho_g(\mathbf{r}_q, t_p)$ denotes the gold standard image at the imaging frame t_p and $\rho(\mathbf{r}_q, t_p)$ denotes the images reconstructed from different navigator sampling patterns at the imaging frame t_p. Figure 4.3 (a) depicts the TAMSE map of an image reconstructed from a spiral navigator. As can be seen in Fig. 4.3 (a), the pixel intensity corresponds to the amount of accumulated error in time. Brighter pixels in the TAMSE map usually reside in the upper vocal tract region of the speech subject's brain. This suggests that articulators in this region are the major source of motion-induced error. In other regions of the speech subject's brain, however, there is seldom any bright pixel. This suggests that the speech subject's brain is less distorted by motion-related noise.



(a) (b)

Figure 4.3: (a) The TAMSE map of an image reconstructed from a spiral navigator and (b) TAMSE map of a noisy image (SNR = 10 dB) reconstructed from the same spiral navigator.

The TAMSE has also been applied to explore the ability of the navigator sampling pattern to capture spatiotemporal dynamics in the presence of noise. Specifically, zero-mean Gaussian white noise is added to the ***k***-space of the phantom data set across all temporal frames under three SNR levels, 5 dB, 10 dB and 15 dB. The TAMSE images can be obtained from reconstructing these noisy phantom data with different navigator sampling patterns. Figure 4.3 (b) depicts the TAMSE map of a noisy image (SNR = 10 dB) reconstructed from the same spiral navigator as in Fig. 4.3 (a). It is obvious from this image that the bright pixels are less obvious in the tongue root and the jaw. This suggests that reconstruction error in the tongue root and the jaw is around the noise level. On the contrary, bright pixels remain obvious around the lips, the tongue tip and the velum. This suggests that reconstruction error in these regions is above the noise level. By comparison with the TAMSE map, it is possible to assess the ability of the navigator sampling pattern to suppress reconstruction error in a specific region.

Conditioning of a problem refers to the sensitivity of a system's output with regard to errors in its input [81]. Mathematically, conditioning is parameterized by the condition number [81]. For a general problem $\mathbf{A}x = \mathbf{b}$, the condition number can be defined in the following way [81],

$$\kappa = \|A^{-1}\|_P \|A\|_P, \tag{4.8}$$

where $\|\cdot\|_P$ denotes the p-norm of a matrix. Usually the condition number is measured with the 2-norm of a matrix, which is defined as the ratio of the largest singular value to the smallest singular value in the singular value decomposition of that matrix [81]. In the basic PS reconstruction, specifically, conditioning is closely related to the $\boldsymbol{\Phi}_q$ matrix [81]. The $\boldsymbol{\Phi}_q$ matrix encompasses the temporal basis functions extracted from the navigator data set and has been previously defined in section 3.4.1 as,

$$\boldsymbol{\Phi}_q = \begin{bmatrix} \varphi_1(t_{1,q}) & \varphi_2(t_{1,q}) & \cdots & \varphi_L(t_{1,q}) \\ \varphi_1(t_{2,q}) & \varphi_2(t_{2,q}) & \cdots & \varphi_L(t_{2,q}) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(t_{P,q}) & \varphi_2(t_{P,q}) & \cdots & \varphi_L(t_{P,q}) \end{bmatrix}. \tag{4.9}$$

The basic PS reconstruction extracts the spatial basis functions by solving the following matrix equation,

$$\boldsymbol{\Phi}_q \boldsymbol{\Psi}_q = \boldsymbol{d}_q, \tag{4.10}$$

where $\boldsymbol{\Psi}_q$ denotes the matrix that encompasses the spatial basis function and $\boldsymbol{d}_q$ denotes the acquired imaging data. The least square solution to the above matrix equation can be obtained through equation (4.11),

$$\boldsymbol{\Psi}_q = (\boldsymbol{\Phi}_q^{\mathrm{H}} \boldsymbol{\Phi}_q)^{-1} \boldsymbol{\Phi}_q^{\mathrm{H}} \boldsymbol{d}_q. \tag{4.11}$$

If we define $\mathbf{A} = \boldsymbol{\Phi}_q$, the condition number of the matrix **A** indicates how PS model-based reconstruction would be degraded by small errors in the acquisition of navigator signals [81]. An ill-conditioned **A** suggests that a small error resulted from sampling of navigator data as well as the subsequent singular value decomposition would cause large deviation from true solution [85]. In the result analysis given later in this chapter, the condition number is used to assess the suitability of the estimated temporal basis for the extraction of spatial coefficients.

The error map is used to qualitatively measure the difference between the gold standard image and the reconstructed image. This qualitative comparison is performed on a pixel-by-pixel

basis. The comparison results indicate the loss of spatial features in reconstruction due to the difference in the navigator sampling patterns. Usually the image difference is small, and the error map is often scaled for better visualization.

The strip plot is used to qualitatively measure how well the reconstruction captures temporal dynamics of speech. In the region of interest, the strip plot provides a panorama of temporal variations, the sharpness of which indicates how well the articulator motion is represented by the temporal basis functions in the PS model [81]. Since the strip can be placed in any location in the region of interest along any arbitrary direction, the strip plot offers a lot of freedom in investigating the temporal events in different vocal tract regions. This property of the strip plot provides a unique perspective to analyze speech dynamics. Figure 4.4 depicts a strip plot of the lip motion across 500 imaging frames.

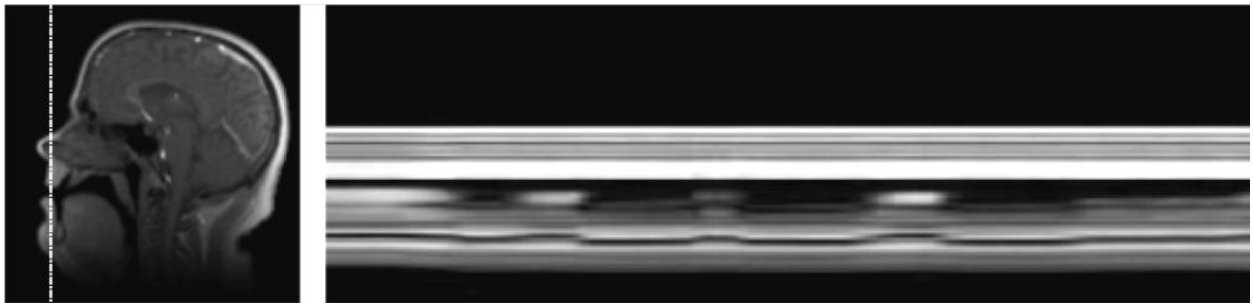


Figure 4.4: A strip plot of the lip motion across 500 imaging frames.

In the above strip plot, the sharpness of the rises and falls of the "speech motion wave" represents the temporal dynamics of the corresponding speech motion. Blurred rises and falls suggest that the articulator motion from one imaging time frame cannot be effectively distinguished from that in another imaging time frame. To fully capture the vocal tract dynamics, this thesis places the strip of pixels mainly in three vocal tract locations, the upper and lower lips, the tongue tip and the velum.

4.1.2.2 Comparison in terms of quantitative and qualitative metrics

In this section, a variety of quantitative and qualitative metrics have been applied to analyze the effects of navigator sampling patterns on the temporal subspace quality. In order to systematically compare these navigator sampling patterns, their ***k***-space trajectories are categorized into five groups as follows:

1. The first group is the ideal Cartesian sampling patterns that use 2, 4, 8, 16 and 24 phase encoding lines for navigating images with a matrix size of 128. These Cartesian patterns are labeled from Cartesian trajectory 1 to Cartesian trajectory 5, or CA1 to CA5 for short.
2. The second group is the conventional Cartesian sampling patterns that use PE lines {63, 64, 65}, {123, 124, 125}, {93, 94, 95}, {3, 63, 123} and {33, 63, 93} for navigation. These Cartesian patterns are labeled from Cartesian trajectory 6 to Cartesian trajectory 10, or CA6 to CA10 for short.
3. The third group is the conventional spiral sampling patterns for navigation. This group includes eight spiral trajectories, five of which rotate the original trajectories with 0 °, 45 °, 90 °, 135 ° and 180 ° rotation angles, three of which have adjusted density in the ***k***-space center. These spiral patterns are labeled from spiral trajectory 1 to spiral trajectory 8, or SP1 to SP8 for short.
4. The fourth group is the conventional radial sampling patterns for navigation. This group includes six radial trajectories that rotate the original trajectory with 0 °, 15 °, 30 °, 45 °, 60 ° and 75 ° rotation angles. These radial patterns are labeled from radial trajectory 1 to radial trajectory 6, or RA1 to RA6 for short.
5. The member sampling patterns for the fifth group depend on the specific metric used for comparison. Member sampling patterns are picked from the second to the fourth group that have the best performance with regard to a specific metric. In other words, this group is set up with an attempt to find an optimized sampling pattern from the above groups.

4.1.2.2.1 Comparison in terms of NRMSE

Comparison of reconstruction quality has been performed in terms of NRMSE on five groups of navigator sampling patterns. Comparison results in terms of NRMSE are shown in Figs. 4.5 (a) – (e). As can be seen with these figures, the NRMSE values between the reconstructed image sequence and the gold standard image sequence are upper bounded by 6.5%. This suggests that overall the basic PS reconstruction can capture the speech dynamics without significant visual loss, regardless of which sampling patterns are used for navigation. Also, difference choices of navigators result in less than 2.5% difference in the NRMSE values. This suggests that difference in the navigator sampling patterns does not yield significant changes in reconstruction. However, slight difference in reconstruction quality exists with respect to each specific navigator chosen.

Specifically, for the ideal Cartesian trajectories (CA1 – CA5), the value of NRMSE decreases when the ***k***-space coverage of the navigator sampling pattern increases. This suggests some temporal basis functions cannot be directly captured at the low-frequency region of the ***k***-space. Rather, these temporal basis functions may reside in higher frequency regions of the ***k***-space. It is beneficial, therefore, to increase the navigator coverage in order to capture more temporal basis functions.

Among other conventional trajectories (CA6 – CA10, SP1 – SP8, RA1 – RA6), some trajectories yield relatively lower NRMSE values in their groups. These trajectories include CA9, CA10, SP1, SP4 and RA5. Given their advantages in suppressing reconstruction error, these trajectories are grouped with the widely used CA6 sampling pattern in Fig. 4.5 (e) for further comparison. Further analysis suggests that SP1 yields the lowest NRMSE among all navigator trajectories. Another spiral trajectory SP4, however, also yields lower error compared to its Cartesian and radial counterparts. Compared with the ideal Cartesian sampling pattern in Fig. 4.5 (a), the performance of this two-turn SP1 spiral trajectory is equivalent to using 16 Cartesian lines for navigation, although SP1 requires less time to traverse the ***k***-space.

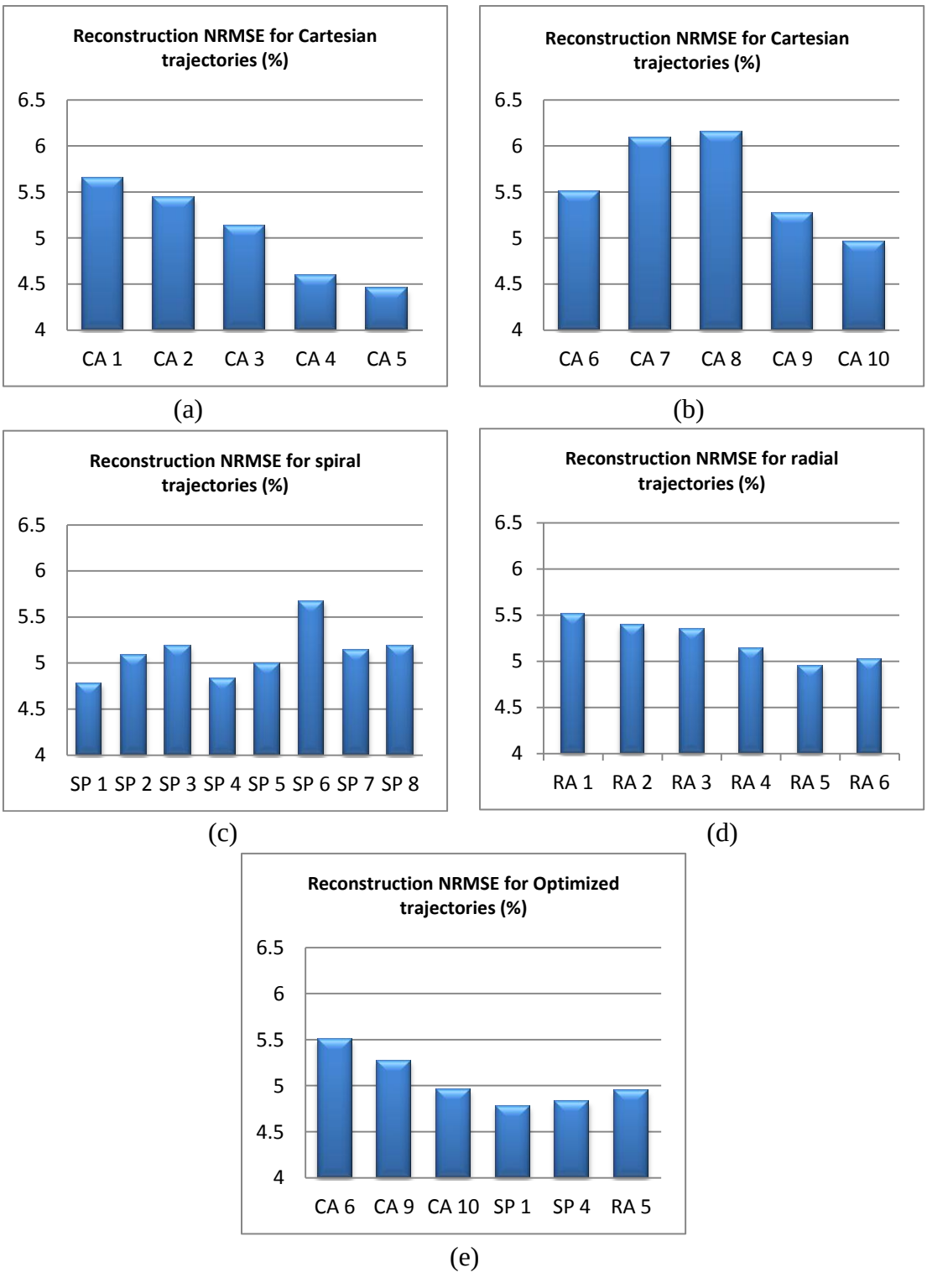


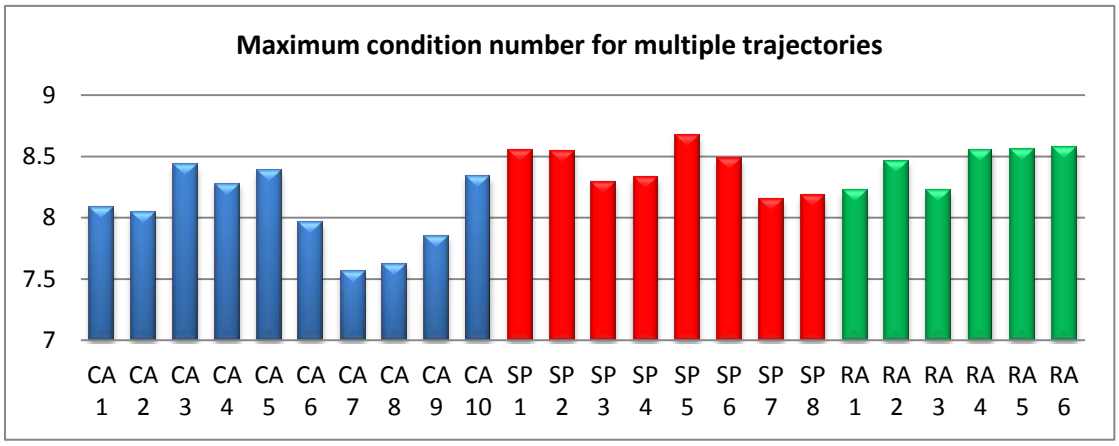
Figure 4.5: Reconstruction NRMSE for (a) the Cartesian trajectories (group 1), (b) the Cartesian trajectories (group 2), (c) the spiral trajectories, (d) the radial trajectories and (e) the optimized trajectories.

4.1.2.2.2 Comparison in terms of conditioning

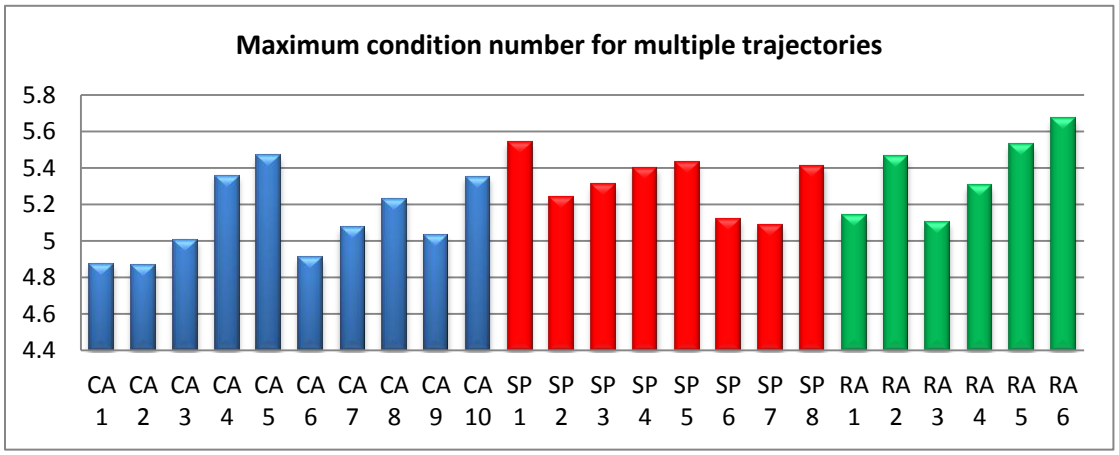
Comparison on the navigator sampling patterns has been performed in terms of the conditioning of the $\boldsymbol{A}$ matrix defined in Section 4.1.2.1. Conditioning of the $\boldsymbol{A}$ matrix reveals the sensitivity of least-square fitting to noise in the basic PS reconstruction. Five groups of navigator sampling patterns are analyzed in terms of the conditioning of the $\boldsymbol{A}$ matrix. For each sampling pattern, conditioning is measured with data sets containing 10 data frames, 40 data frames and 100 data frames, where a data frame refers to 128 consecutive imaging frames.

To compactly demonstrate the results, the maximum condition number among all rows of the $\boldsymbol{A}$ matrix is picked out for comparison. The maximum condition number provides a worst case in which reconstruction quality deteriorates along with small perturbations in the navigator sampling errors. Specifically, Figs. 4.6 (a) - (c) depict the maximum condition number for 10-frame, 40-frame and 100-frame data.

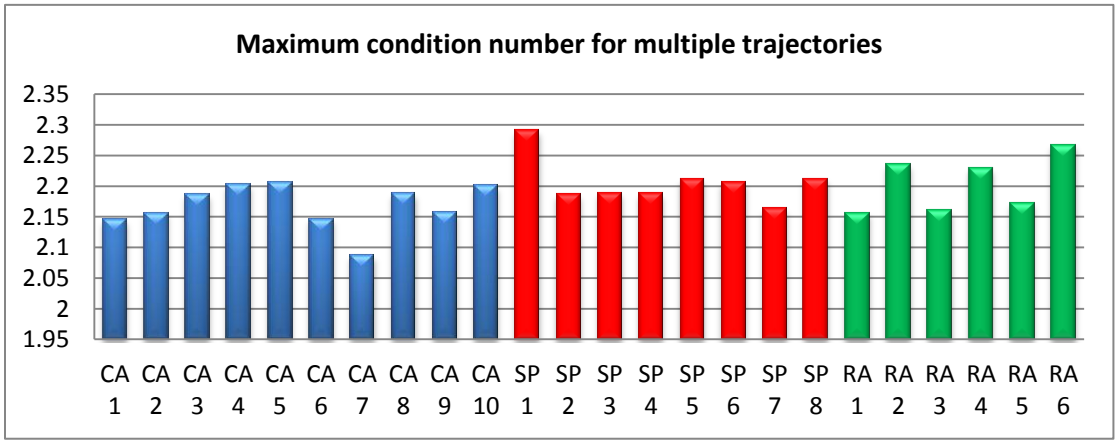
Three observations can be made from Fig. 4.6. First, overall the maximum condition number decreases as the number of data frames used for navigation increases. This suggests the PS model-based reconstruction is more robust as the amount of acquired data increases. Secondly, there exists no significant difference between the maximum condition numbers for Cartesian, spiral or radial navigator trajectories. Thirdly, the values of the maximum condition numbers suggest that no particularly ill-conditioned $\boldsymbol{A}$ matrices exist. Specifically, the maximum condition numbers range around 8, 5 and 2 when 10-frame, 40-frame and 100-frame data sets are used for navigation. The values of the condition numbers suggest that data fitting in the PS model for the speech data set is not significantly affected by the particularly chosen navigator trajectory.



(a)



(b)



(c)

Figure 4.6: Analysis on the maximum condition number of the $\boldsymbol{A}$ matrix for multiple navigator trajectories using a (a) 10-data-frame speech phantom, (b) 40-data-frame speech phantom and (c) 100-data-frame speech phantom.

4.1.2.2.3 Comparison in terms of basic PS reconstructions

Reconstruction images and their corresponding error maps serve as powerful means to evaluate the spatiotemporal resolution of the reconstructed images. Similarly, with the previous analyses, reconstruction and error maps have been examined for five groups of navigator sampling patterns. For each sampling pattern, representative imaging frames that reflect important vocal tract shaping were chosen from the reconstructed image sequence. Although extensive comparisons have been performed, only CA 6, CA 9, CA 10, SP 1, SP 4 and RA 5 are shown in this thesis for simplicity. Specifically, Fig. 4.7 depicts localized differences among the reconstructed images in terms of the velum and the tongue tip. Figure 4.8 depicts localized differences among the error maps in terms of the tongue tip and the tongue dorsum. Image errors in the error maps are scaled by a factor of 20 for better visualization.

Overall, reconstructions from multiple navigator data sets can capture major articulator shaping and speech dynamics. There exists no significant difference in terms of reconstruction quality. However, differences among the reconstructed images and the error maps can be observed in the localized features of small-size articulators, such as the end of the velum, the tongue root and the tongue tip. These localized differences are indicated with white arrows. From Fig. 4.7, SP 1 has better performance in capturing the shape of the velum, i.e., the small crescent shaped velum end can be effectively differentiated from the tongue root; both SP 1 and RA 6 better visualize curvature of the tongue tip as well as the air stream between the tongue tip and the palate. From Fig. 4.8, SP 1, SP 4 and RA 5 have better performance in suppressing the image errors in the velum; SP 1 and SP 4 can better capture the speech dynamics of the tongue root. Although SP 4 shows lower image error in the tongue root, SP 1 has better overall performance in terms of image errors. To conclude, SP 1 has slightly better performance in capturing the articulator shaping in the human vocal tract.

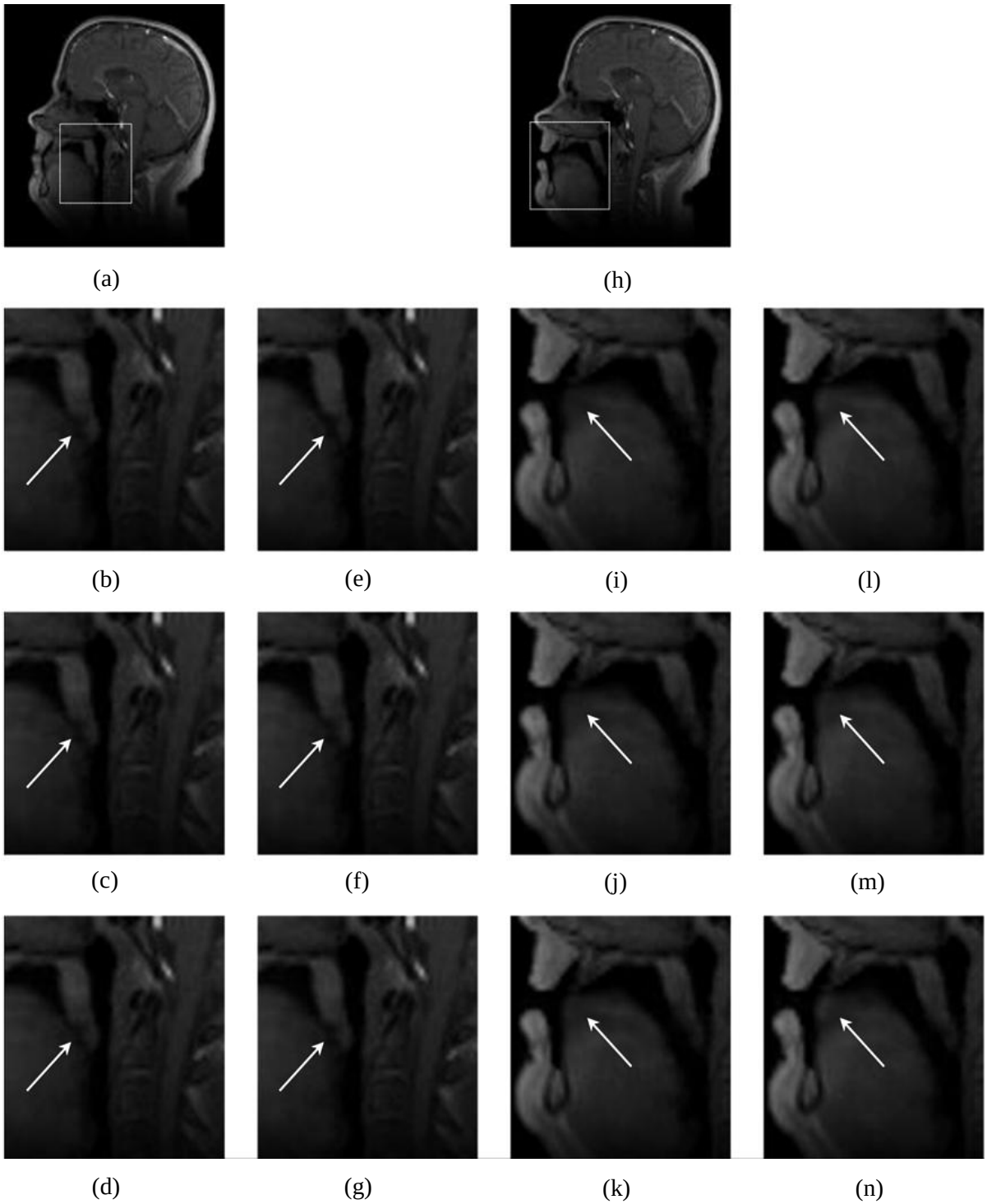


Figure 4.7: Reconstructed images show difference in the velum (a) - (g) and the tongue tip (h) - (n). Gold standard: (a), (h). Cartesian trajectory 6: (b), (i); Cartesian trajectory 9: (c), (j). Cartesian trajectory 10: (d), (k). Spiral trajectory 1: (e), (l). Spiral trajectory 4: (f), (m). Radial trajectory 5: (g), (n).

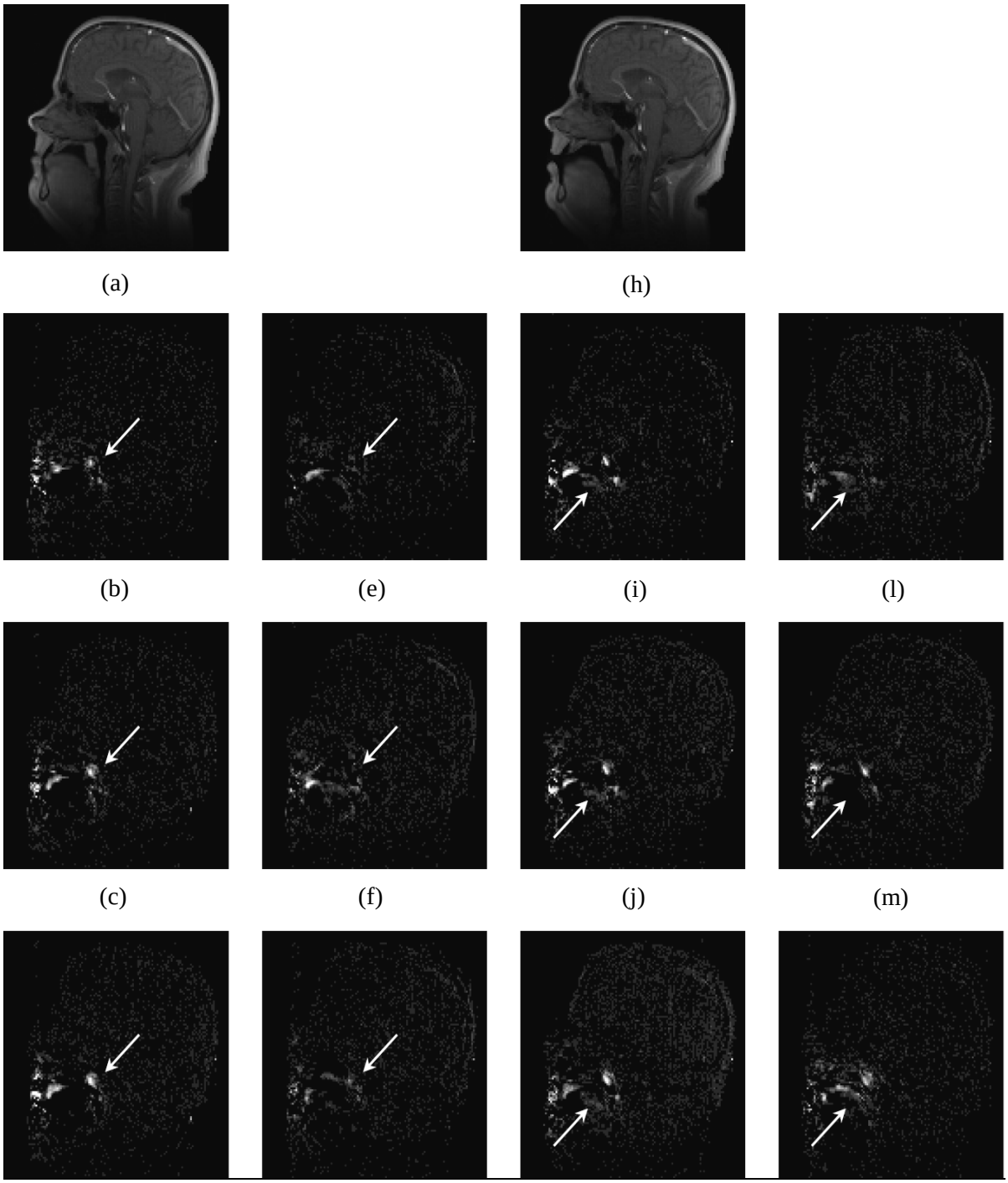


Figure 4.8: Reconstructed images show difference in the velum (a) - (g) and the tongue root (h) - (n). Gold standard: (a), (h). Cartesian trajectory 6: (b), (i); Cartesian trajectory 9: (c), (j). Cartesian trajectory 10: (d), (k). Spiral trajectory 1: (e), (l). Spiral trajectory 4: (f), (m). Radial trajectory 5: (g), (n).

4.1.2.2.4 Comparison in terms of TAMSE

Comparison on the navigator sampling patterns has been performed in terms of TAMSE. Analysis on TAMSE has been performed on a noise-free speech data set with the above-mentioned five groups of navigator sampling patterns. In addition, the TAMSE for these sampling patterns is also analyzed on noisy speech data sets with three different noise levels, 5 dB, 10 dB and 15 dB. Although extensive comparisons have been performed, only three sets of representative results are shown to simplify discussion. Specifically, Figs. 4.9 (a) – (e) depict the TAMSE maps for the ideal Cartesian navigators on a noise-free speech data set. Figures 4.10 (a) – (f) depict the TAMSE maps for the optimized conventional navigators on a noise free speech data set. Figures 4.11 (a) – (f) depict The TAMSE maps for the optimized conventional navigators on a noisy speech data set with an SNR of 10 dB.

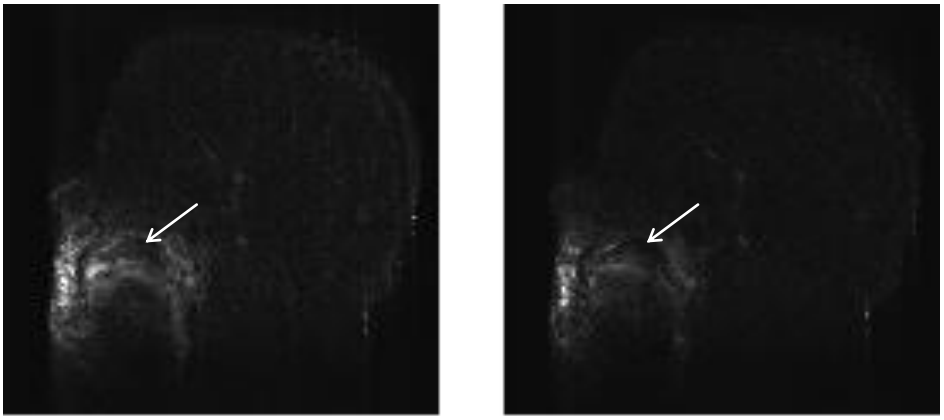
As can be seen in these figures, the average reconstruction error between the gold standard and the reconstructed image sequence mainly resides in the upper vocal tract. Error is more likely to accumulate in regions where the shape of soft-tissue structures quickly varies. These dynamic regions include the jaw, the upper lip, the lower lips, the tongue tip, the tongue root as well as the velum. The TAMSE maps effectively demonstrate the navigators' ability to capture fast-varying speech dynamics.

Let us first consider the case of ideal Cartesian navigator trajectories in a noise-free data set. From Figs. 4.9 (a) – (e), error intensity gradually decreases as the number of phase encoding lines increases in the navigator sampling pattern. Comparing Figs. 4.9 (d) – (e) with Figs. 4.9 (a) – (c), it is obvious that the TAMSE value in the vocal cavity (where the arrow points) is suppressed when 16 or more Cartesian phase encoding lines are used for navigation. A larger number of phase encoding lines leads to decreased TAMSE intensity, which usually suggests that more temporal features are preserved in the entire PS model-based reconstruction.

Consider the case of optimized navigator trajectories in a noise free data set. As can be seen with Figs. 4.10 (a) – (f), the optimized navigator trajectories have similar overall performance,

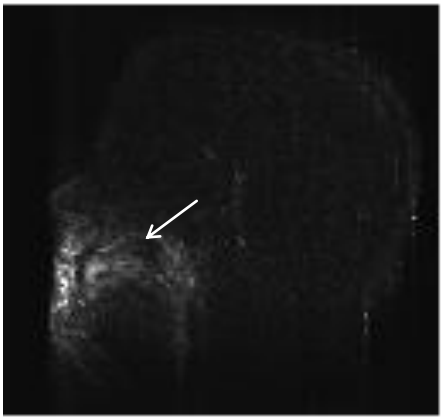
i.e., no trajectory has significantly lower TAMSE value than other trajectories in the group. However, SP 4 has slightly lower error in the oral cavity and the upper and lower lips (as indicated with the white arrows).

Consider the case of optimized navigator trajectories in a noisy data set with an SNR of 10 dB. As can be seen in Figs. 4.11 (a) – (f), the noisy data set causes the average error to spread out across the vocal tract region. Especially when compared with Fig. 4.10 (a) – (f), a clear "stripe-shaped" error pattern is shown in the horizontal direction of lip motion in all optimized navigator trajectories. Among these trajectories, however, CA 6 has better performance in reducing the structural average error in the vocal tract. As can be seen in the image, Fig. 4.11 (a) has reduced error in the tongue dorsum and the velum. This phenomenon can be explained from the fact that CA 6 samples only the center of ***k***-space, which has higher SNR than other ***k***-space regions. These observations suggest that CA 6 may bring potential benefit to noisy speech applications.

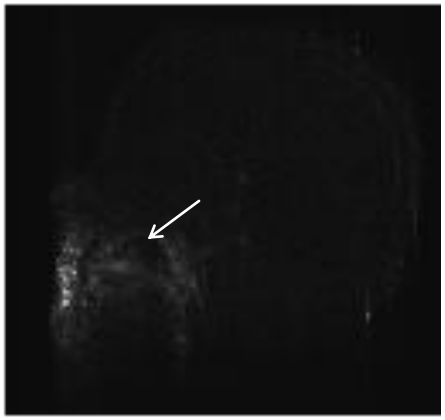


(a)

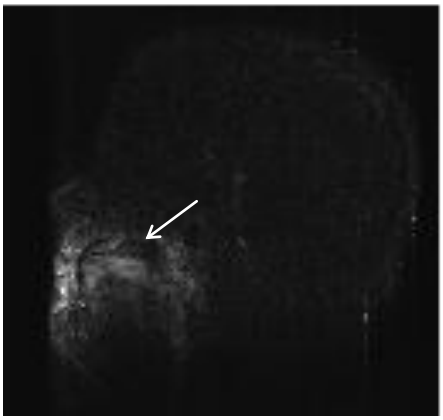
(d)



(b)



(e)



(c)

Figure 4.9: Reconstruction TAMSE for the ideal Cartesian trajectories: (a) Cartesian trajectory 1, (b) Cartesian trajectory 2, (c) Cartesian trajectory 3, (d) Cartesian trajectory 4 and (e) Cartesian trajectory 5.

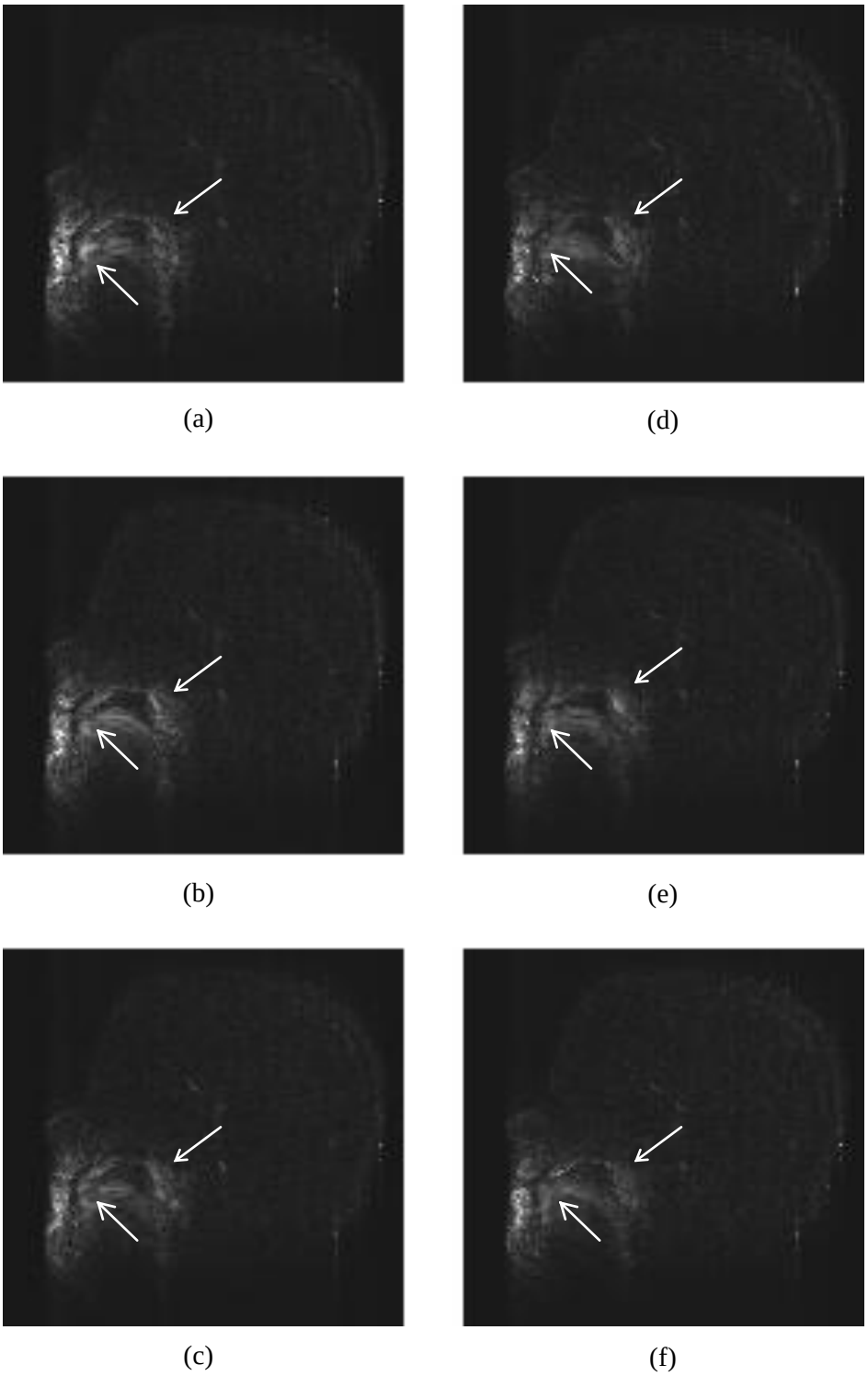


Figure 4.10: Reconstruction TAMSE for the optimized trajectories: (a) Cartesian trajectory 6, (b) Cartesian trajectory 9, (c) Cartesian trajectory 10, (d) spiral trajectory 1, (e) spiral trajectory 4, (f) radial trajectory 5.

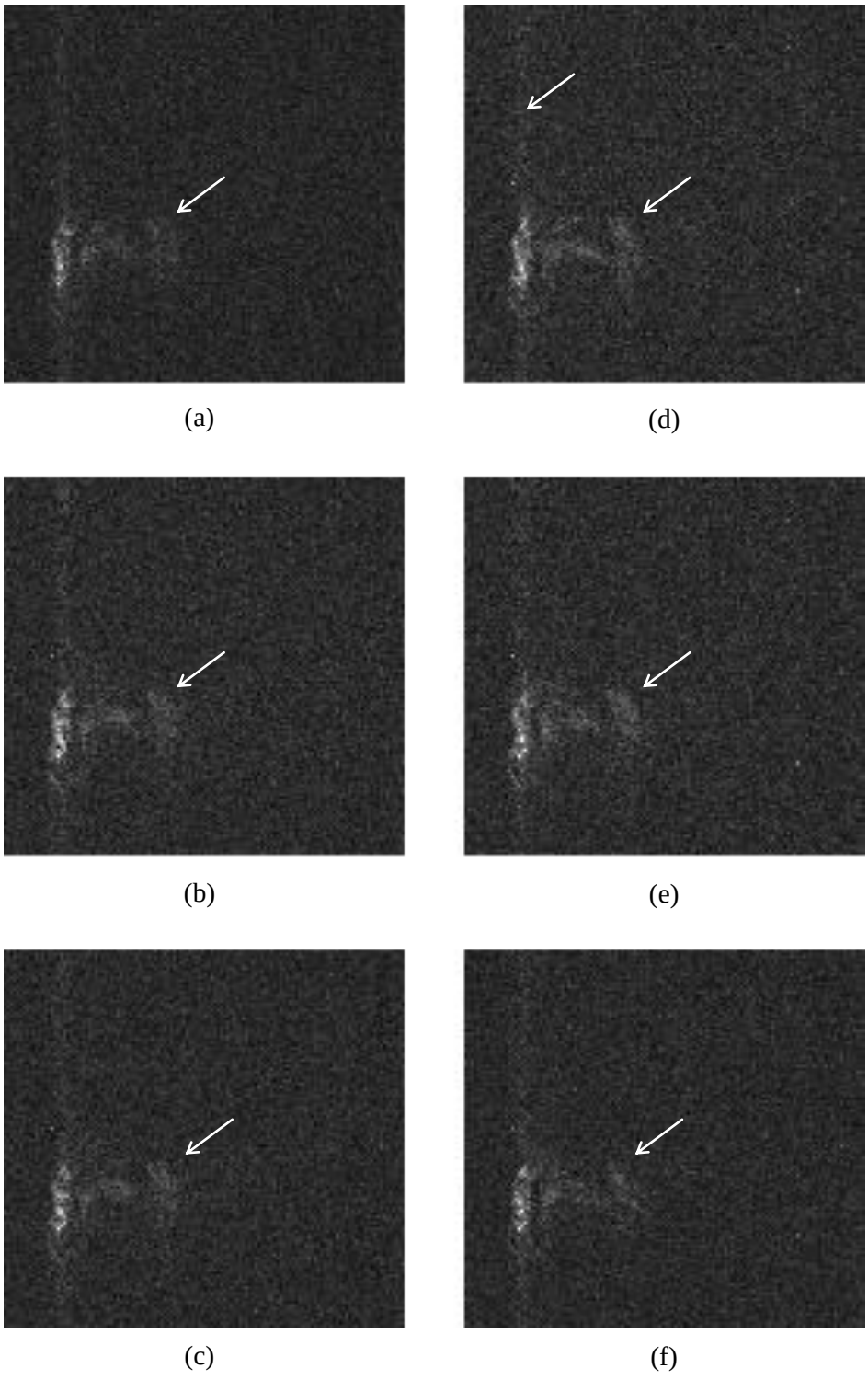


Figure 4.11: Reconstruction TAMSE from a noisy data set: (a) Cartesian trajectory 6, (b) Cartesian trajectory 9, (c) Cartesian trajectory 10, (d) spiral trajectory 1, (e) spiral trajectory 4, (f) radial trajectory 5.

4.1.2.2.5 Comparison in terms of strip plot

Comparisons of the navigator sampling patterns have been performed in terms of the strip plot in the *(**x**, t)*-space. Similarly with the analysis of TAMSE, the strip plot has been used to analyze five groups of navigator sampling patterns. Although extensive comparisons have been performed, only CA 6, CA 9, CA 10, SP 1, SP 4 and RA 5 are chosen for discussion. For each sampling pattern, three speech locations in the vocal tract are selected for strip plot: the lips, the tongue tip and the velum. The temporal dynamics of these regions are of special interest in phonetic and acoustic research. Specifically, Fig. 4.12 depicts the strip plot of the upper and lower lips; Fig. 4.13 depicts the strip plot of the tongue tip; Fig. 4.14 depicts the strip plot of the velum. The locations of the strips of pixels are depicted with bold dotted lines.

From the strip plots, it is obvious that basic PS reconstruction can capture the overall temporal transitions between different speech events regardless of the specific ***k***-space trajectory chosen for navigation. Comparing the strip plots of the reconstructed images with the gold standard strip plots in Figs. 4.12 (b), 4.13 (b) and 4.14 (b), overall movements of the upper and lower lips, the tongue and the velum are well captured with all ***k***-space trajectories. However, the performance of navigator trajectories varies in the abilities to resettle localized temporal dynamics. Generally speaking, the non-Cartesian trajectories provide better temporal dynamics than their Cartesian counterparts. For instance, Fig. 4.12 (c) has less blurring than (f) between two successive imaging frames. Sharper images of temporal transition of the closed lips and the open lips are captured by SP 1 than by the CA 6. Similar results can be seen with the error maps. The error map in Fig. 4.13 (d) and (f) demonstrates that CA 9 has "brighter blurring" than in SP 1. This blurring corresponds to the contact between the tongue tip and the hard palate and is of significance to phonetic research. From these observations, non-Cartesian navigator trajectories are relatively more beneficial for preserving the temporal transitions in speech. This feature is desirable for applications that require high temporal resolution.

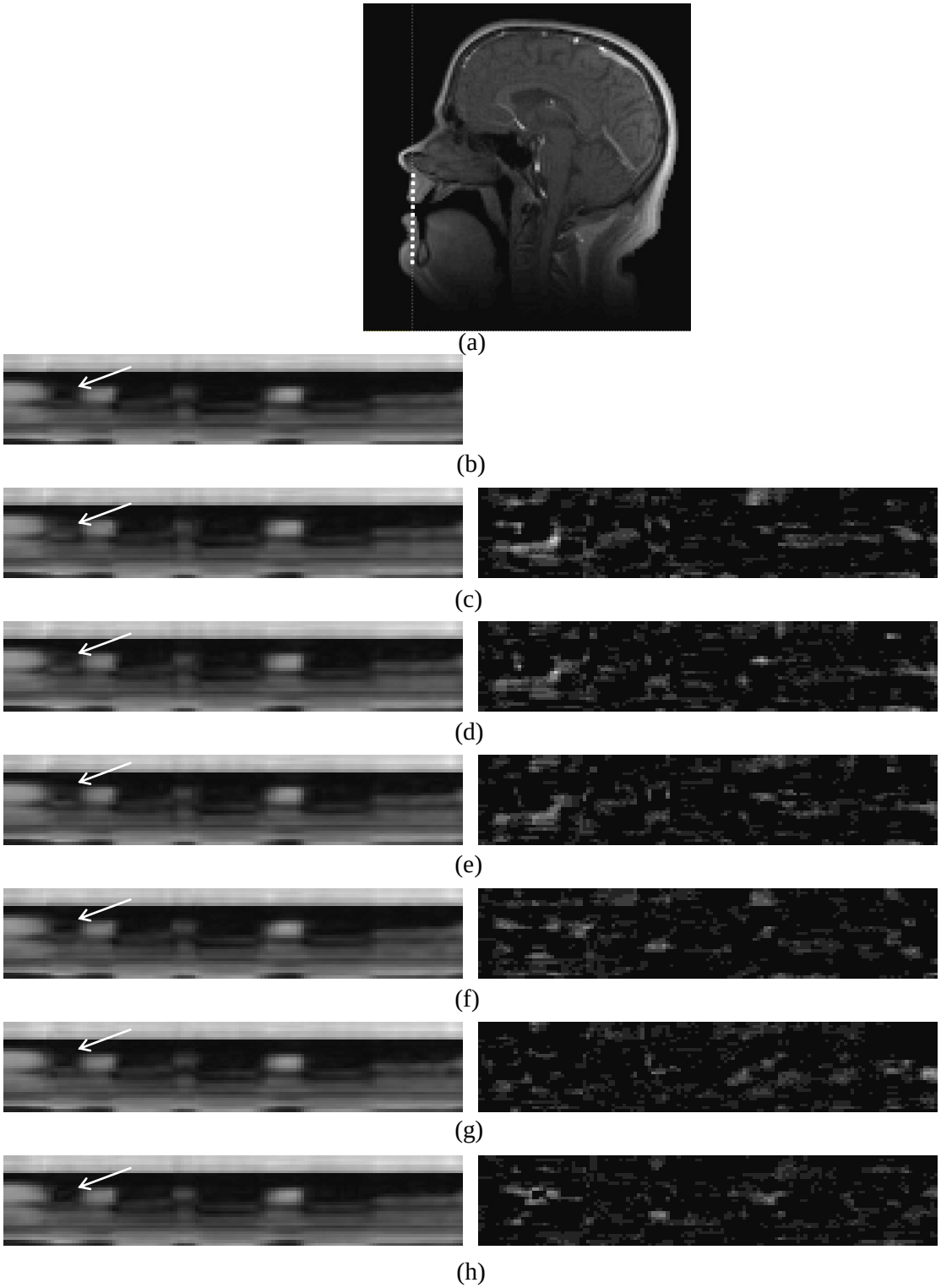


Figure 4.12: Strip plot of temporal dynamics for the upper lip. (a) Indication of strip plot location, (b) gold standard, (c) Cartesian trajectory 6, (d) Cartesian trajectory 9, (e) Cartesian trajectory 10, (f) spiral trajectory 1, (g) spiral trajectory 4, (h) radial trajectory 5.

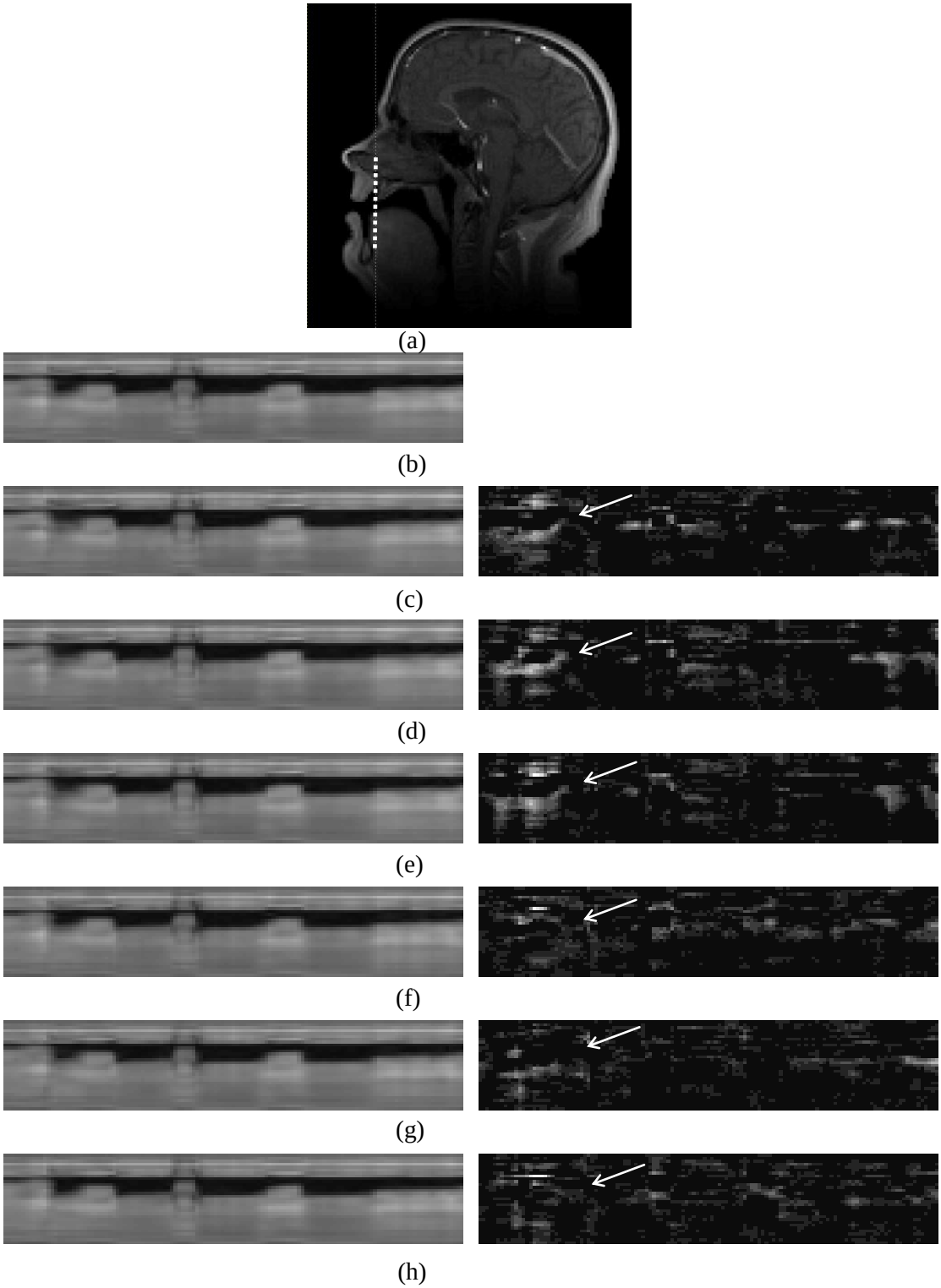


Figure 4.13: Strip plot of temporal dynamics for the tongue tip. (a) Indication of strip plot location, (b) gold standard, (c) Cartesian trajectory 6, (d) Cartesian trajectory 9, (e) Cartesian trajectory 10, (f) spiral trajectory 1, (g) spiral trajectory 4, (h) radial trajectory 5.

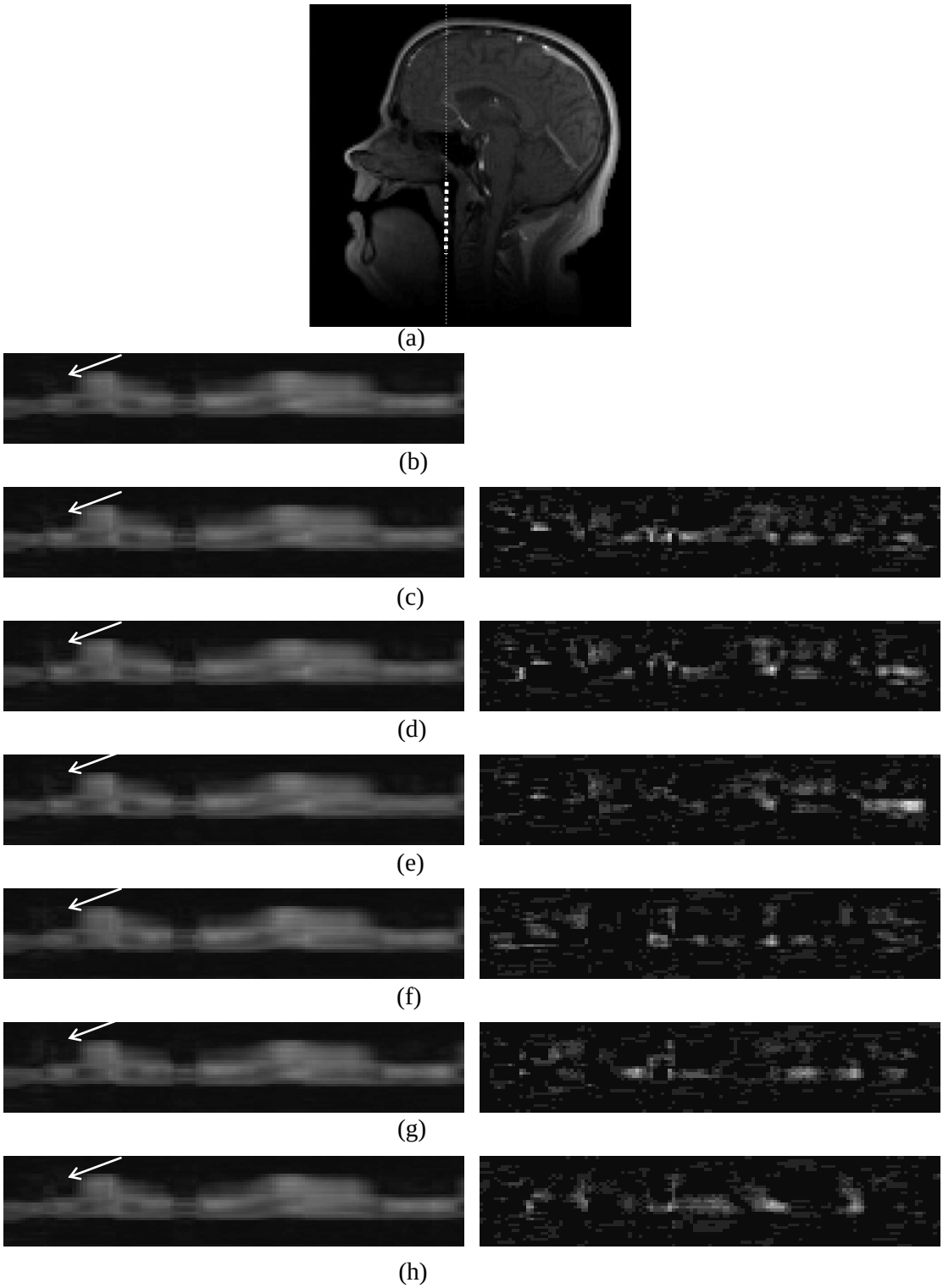


Figure 4.14: Strip plot of temporal dynamics for the velum. (a) Indication of strip plot location, (b) gold standard, (c) Cartesian trajectory 6, (d) Cartesian trajectory 9, (e) Cartesian trajectory 10, (f) spiral trajectory 1, (g) spiral trajectory 4, (h) radial trajectory 5.

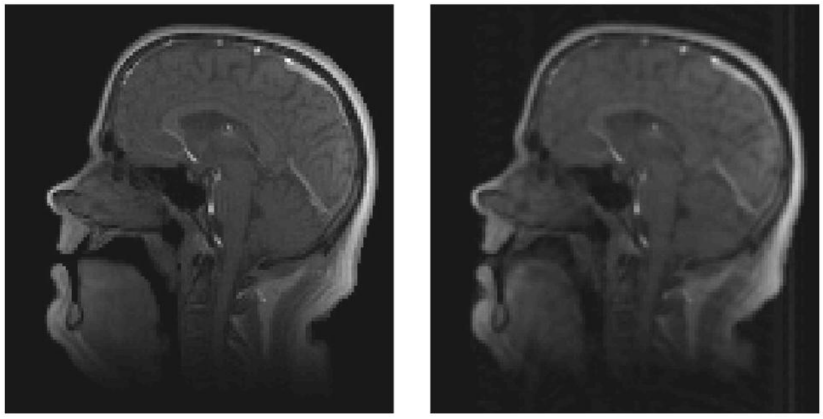
4.1.3 Comparison in terms of PS model-based reconstruction and alternative methods

4.1.3.1 Basic PS reconstruction and sliding window reconstruction

The sliding window method is commonly used to reconstruct dynamic image sequences from the sparsely sampled data. As a basic reconstruction approach, it has been combined with various sampling patterns to perform dynamic imaging experiments [98]. The principle for sliding window reconstruction is straightforward: accumulated imaging data can be obtained from sliding a ***k***-space acquisition window over multiple imaging time frames [98]. Without loss of generality, let us consider the case of time-sequential sampling for Cartesian trajectories although other more efficient sampling schemes could be used in practice. Generally in time-sequential Cartesian sampling, only one phase encoding line can be acquired within one imaging time frame. However, the sliding window method, as its name suggests, designs a window to hold imaging data acquired across multiple imaging time frames [98]. Within a window, data across multiple previous time frames are considered to be simultaneously acquired with the current data [98]. In this way, data are shared within the acquisition window before they are Fourier transformed to yield dynamic image sequences [98]. The sliding window method, by its nature, attempts to synthesize a larger amount of data from view sharing [98]. Usually the greater amount of data sharing suggests that a larger window size is needed.

Based on the above discussion, it is obvious that sliding window reconstruction does not provide sufficient temporal resolution for dynamic imaging applications. This drawback prevents sliding window reconstruction from recovering true temporal dynamics of the signal. Instead, the temporal features in the reconstructed image sequence can be regarded as an “averaged” temporal variation within the window step size [98]. In addition, the loss in temporal resolution can be fully characterized by the value of the window step size. Moreover, the temporal Nyquist sam-

pling criterion may be violated when a large window step size is used [98]. Although sliding window reconstruction suffers from low temporal resolution and other drawbacks, it is widely used due to its implementation simplicity and computation efficiency [98]. In this thesis, sliding window reconstruction is implemented to reconstruct a dynamic speech imaging data set created from the speech phantom. As a comparison, the basic PS algorithm is also used to reconstruct the same data set. The window step size of the sliding window method is defined as $8T_R$, where T_R denotes the time to acquire one phase encoding line in the Cartesian sampling pattern. A snapshot of the reconstruction result can be seen with Fig. 4.15.



(a) (b)

Figure 4.15: Comparison of basic PS and sliding window reconstruction: (a) basic PS reconstruction and (b) sliding window reconstruction.

From Fig. 4.15, it is obvious that sliding window reconstruction yields a high level of motion artifacts in the reconstructed image. Motion artifacts are especially obvious around the upper and lower lip, the tongue tip, the tongue root and the velum. Articulator shaping in sliding window reconstruction is degraded to a level that the lips, the tongue and the palate seem to be attached, whereas they can be clearly distinguished in the basic PS reconstruction. Also, it is worth noting that choosing window step size as $8T_R$ has already violated the temporal Nyquist sampling criterion. Given the reconstruction result, it is obvious that acceptable image quality is not possible when conventional sliding window reconstruction is used for dynamic speech imaging.

4.1.3.2 Basic PS, basic sparse and PS-sparse reconstruction

The previous section has shown that the sliding window reconstruction would have difficulty to capture speech dynamics due to severe motion artifacts. In this section, instead, we focus on comparing the performance of advanced reconstruction schemes on the speech phantom data set. Specifically, the basic PS reconstruction, the basic sparse reconstruction and the PS-sparse reconstruction are chosen to reconstruct a complex-valued 25-data-frame speech phantom data set of normal speech production. For both the basic PS and the PS-sparse reconstruction, the model order and the regularization parameters are chosen based on empirical evaluation of reconstruction quality. Specifically, a model order of 20 is chosen for both PS model-based reconstruction schemes. Reconstructions are compared in terms of the reconstructed images and the strip plot. Figures 4.16 (a), (b), (c) and (d) depicts the gold standard, the PS reconstruction, the basic sparse reconstruction and the PS-sparse reconstruction respectively.

From the reconstructions shown in Fig. 4.16, overall, three image reconstruction strategies recover localized articulator dynamics of speech production. Among these reconstructions, however, the PS-sparse reconstruction yields better image quality. Specifically, the basic PS reconstruction suffers from slight motion artifacts around fast-varying articulators, such as the tongue, the velum and the lips (as indicated with arrows). These motion artifacts are resulted from ill-conditioning since a high model order of 20 is used for reconstruction. However, these artifacts are successfully suppressed by imposing $(\boldsymbol{x}, f)$-sparsity in PS-sparse. The basic sparse reconstruction is also inferior to PS-sparse reconstruction in capturing motion in the case of limited measured data (25 data frames): the tongue tip and the lower and upper lips are blurred (as indicated with arrows). In addition, the strip plots of both basic PS and basic sparse reconstructions show blurred temporal dynamics during transitions of some speech motion. To conclude, the PS-sparse reconstruction yields refined results by mutually imposing the partial separability and the $(\boldsymbol{x}, f)$-sparsity constraints.

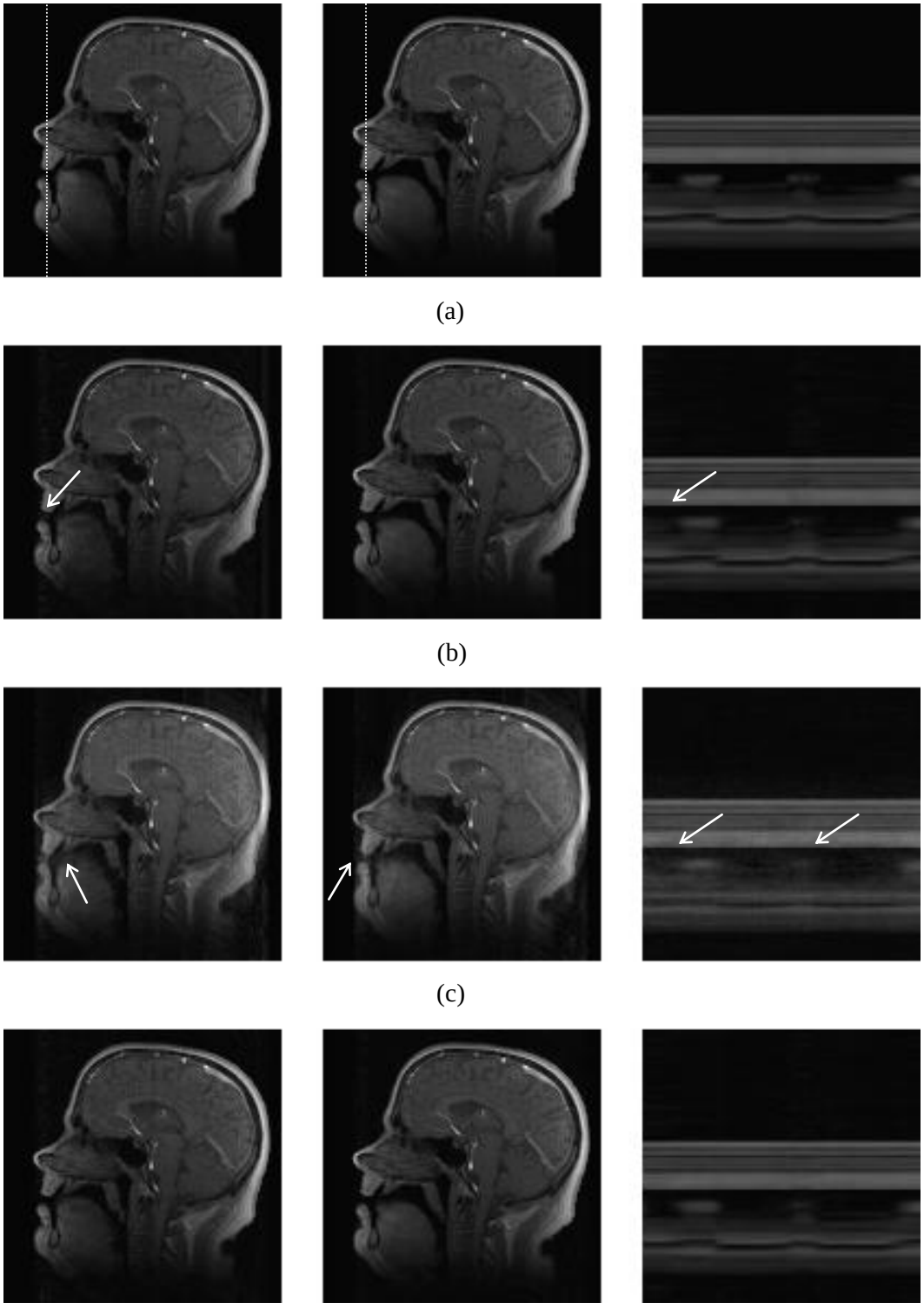


Figure 4.16: Reconstructed frames for two speech movements and strip plot of the tongue tip. (a) The gold standard (dotted lines indicate location of the strip), (b) the basic PS reconstruction, (c) the basic sparse reconstruction, (d) the PS-sparse reconstruction.

4.2 Experiments

4.2.1 Comparison in terms of basic PS and PS-sparse reconstruction

Previous section has demonstrated through a numerical phantom that the PS-sparse method enables high-quality reconstructions with a large model order L and reduced amount of measured data. In this section, we focus on comparing the performance of basic PS-sparse reconstruction and basic PS reconstruction through an *in vivo* speech experiment.

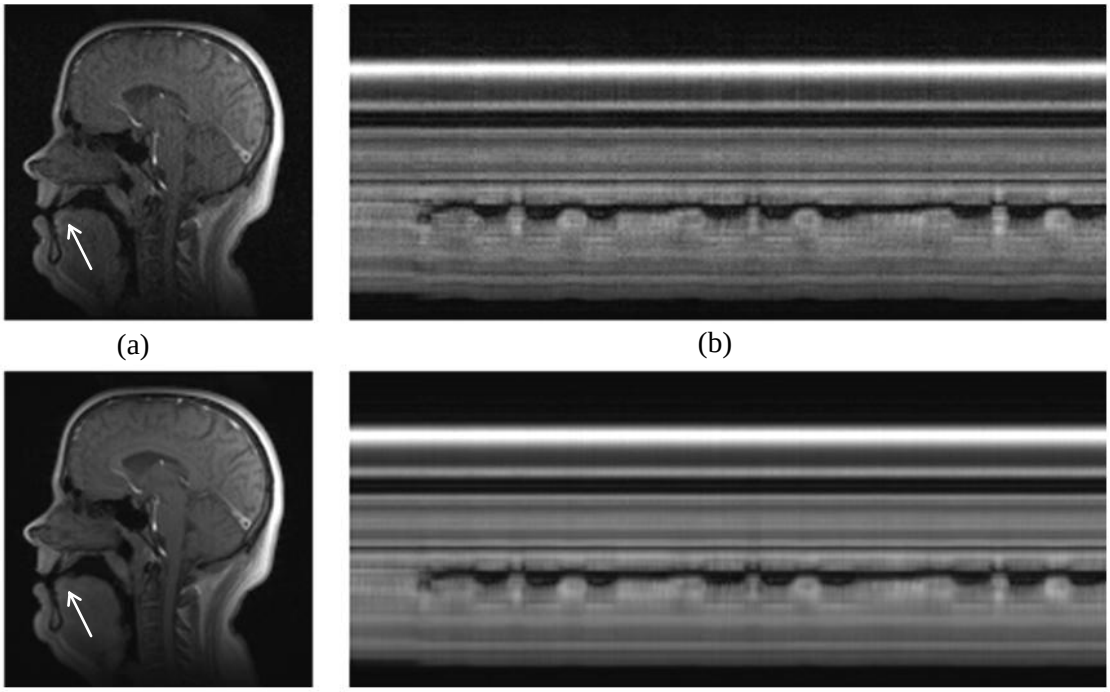
The PS model-based data acquisition scheme was performed on a Siemens Trio 3T scanner with a 12-channel head receiver coil. Repetitive articulator motion during natural speech production of /za/-/na/-/za/ sounds was acquired from one native English speaker at normal speaking speed. The acquired data set covers a 280 mm × 280 mm × 32.5 mm FOV in a single mid-sagittal slice of the upper vocal tract. A spiral trajectory is used to acquire a navigator data set and a Cartesian trajectory is used to acquire an imaging data set. The fast low-angle shot (FLASH) sequence was applied to perform the above imaging experiments in accordance with local internal review boards.

The sampled data from the above *in vivo* speech experiment have in total 200 data frames (25600 imaging frames). In order to systematically explore the performance of PS and PS-sparse, the sampled data were truncated into different data lengths and were reconstructed with different model orders. Specifically, the sampled data were truncated into six data sets consisting of 20 data frames, 40 data frames, 60 data frames, 80 data frames, 120 data frames, 160 data frames and 200 data frames, respectively. These data sets were reconstructed with four model orders of 6, 12, 20 and 30, respectively, to evaluate changes in reconstruction quality. Although systematic explorations have been made, a representative reconstruction (40-data-frame, model order 20) is shown to demonstrate articulator motion in a mid-sagittal slice of the upper vocal tract.

Figures 4.17 (a) and (c) depict the reconstruction of articulator motion during the production

of the /za/ sound based on PS and PS-sparse, respectively. Overall, both PS and PS-sparse can capture dynamic articulator motion in high spatial resolution. Major vocal articulators, such as the upper and lower lips, the tongue tip, the hard palate, the velum and the epiglottis, can be effectively identified from their background. Moreover, white arrows in Figs. 4.17 (a) and (c) indicate that the air stream between the tongue tip and the palate was well captured. However, Fig. 4.17 (b) is less degraded by amplified noise and motion artifacts compared with Fig. 4.17 (a). In basic PS reconstruction, motion artifacts occur mainly in the vicinity of major articulators, such as the velum and the lips, while the amplified noise spreads out across the entire image. Figures 4.17 (c) and (d) depicts the strip plot for basic PS reconstruction and PS-sparse reconstruction, respectively. The strip of pixels is taken through the upper and lower lips (as indicated with dotted lines in Fig. 4.17 (a) and (c)) across 400 consecutive imaging frames. The strip plots indicate that major transitions in the lip motion is captured by both basic PS reconstruction and PS-sparse recontruction. However, overall PS-sparse reconstruction in Fig. 4.17 (d) provides a smoother strip plot than what basic PS reconstruction can offer in Fig. 4.17 (c). This indicates the fact that ill-conditioning issues also degrade the temporal dynamics of basic PS reconstruction. Temporal dynamics of the lip movements can be refined with PS-sparse reconstruction.

The above results are based on the reconstruction of a short-length data set using high model order. As can be seen from the reconstructed images, overall, the articulator motion and speech dynamics can be well reconstructed with both PS and PS-sparse. When an appropriate model order is chosen to reconstruct a data set with sufficient data, PS is a favorable choice since it is straightforward to implement. When a high model order is needed to capture the fine details of articulator motion and temporal events from limited measured data, PS-sparse serves as a better choice due to its ability to suppress ill-conditioning. However, PS-sparse is computationally intensive and requires long reconstruction time.



(a) (b)

(c) (d)

Figure 4.17: (a) Basic PS reconstruction, (b) a strip plot for basic PS reconstruction. (c) PS-sparse reconstruction and (d) a strip plot for PS-sparse reconstruction.

4.2.2 Multislice dynamic speech imaging of vocal tract shaping

The previous section explores the ability of basic PS reconstruction and PS-sparse reconstruction to capture speech dynamics in one imaging plane. However, effective speech analyses usually require reconstruction of articulator dynamics in multiple imaging planes or even from an entire three-dimensional volume. This section focuses on applying PS-sparse to reconstruct experimental data over 5 imaging planes.

The basic experimental setup was similar to that in the previous section. However, the acquired experimental data covered a stack of five mid-saggital imaging slices. Composite spiral navigator trajectory / Cartesian imaging data trajectory were used to acquire data over multiple slices. This sampling pattern is shown in Fig. 4.18, where blue Cartesian lines indicate acquisition of the imaging data and red spiral trajectories indicate acquisition of navigator data.

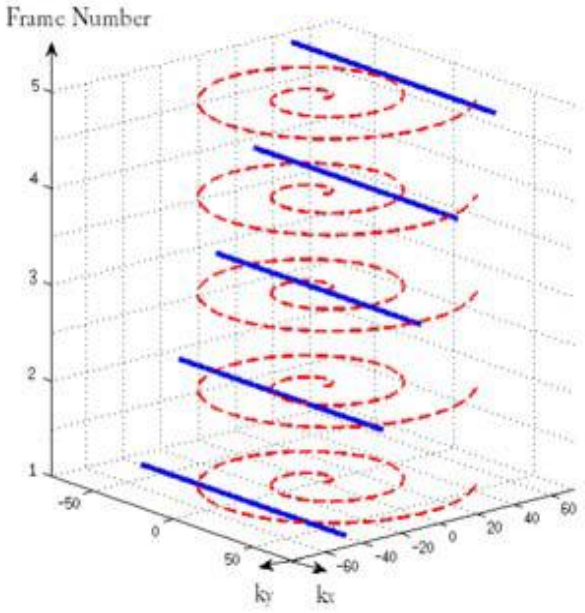


Fig. 4.18: The ***(k**, t)*-space sampling pattern for the dynamic speech imaging experiment.

Reconstruction of the sampled data was performed with PS-sparse at a frame rate of 20 fps for each imaging slice. The model order and regularization parameters for the PS-sparse model were chosen based on the empirical evaluations of reconstruction quality in the previous section. Specifically, a model order of 20 is chosen to recover speech dynamics over multiple slices. Figure 4.19 depicts five mid-sagittal slices of the soft-tissue structures in the oropharyngeal region during normal speech production of the /na/ sound. The reconstructed images of all slices display clear vocal tract shaping and are free of severe motion artifacts or visual losses. The relative positions of major articulators, such as the tongue tip, the palate and the velum, are well captured by Figs. 4.19 (b), (c) and (d) while the contact between the tongue and the palate is not observed in Figs. 4.19 (a) and (e). Based on these results, it is reasonable to infer the natural articulator shaping during the production of the /na/ sound. Specifically, the tongue stays closer to the palate in its middle lines while it stays farther from the palate on both of its edges. If considered from the coronal angle, the ensemble of these imaging slices should display obvious convexity. This phenomenon is known in speech research as “tongue grooving”.

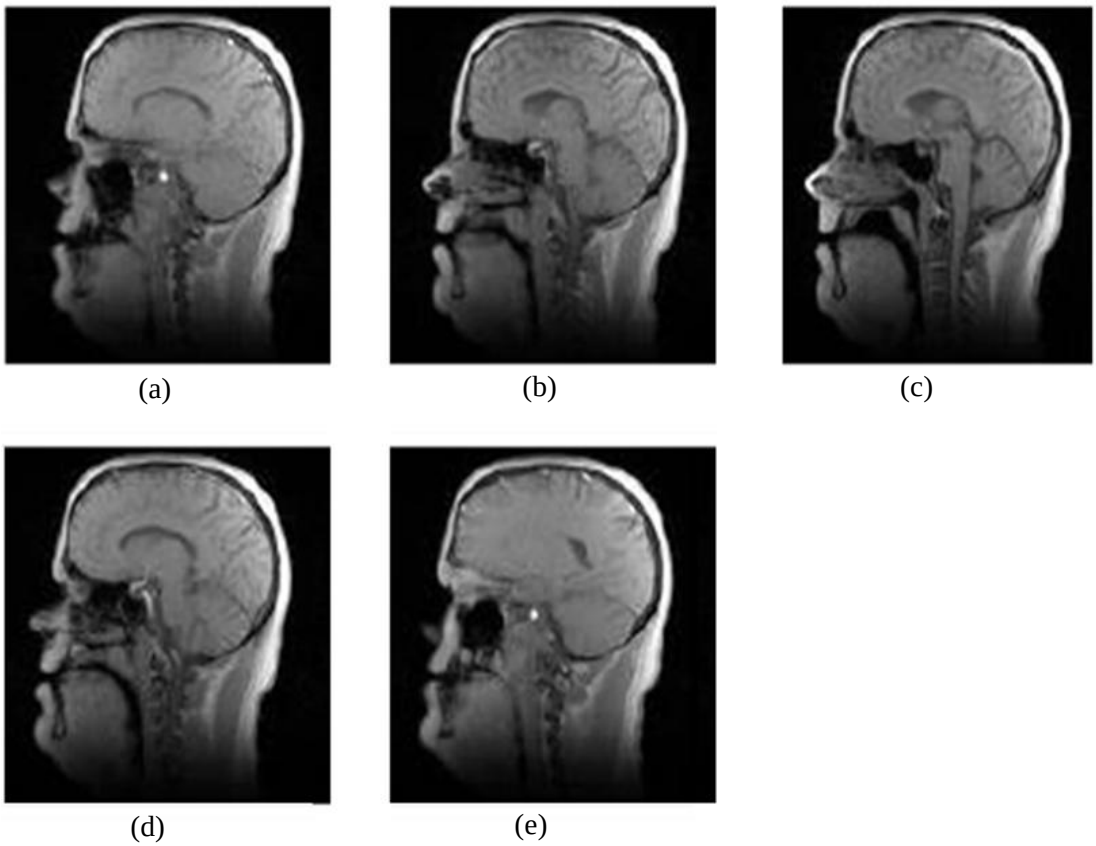
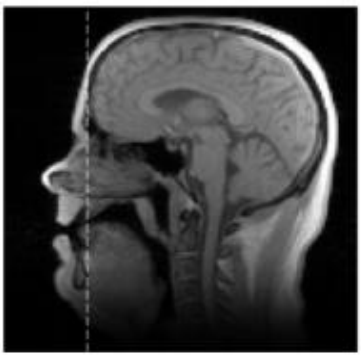
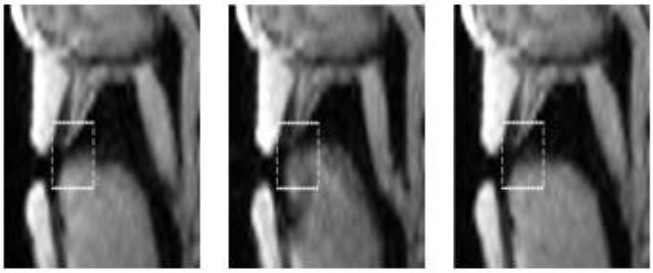


Figure 4.19: Reconstruction of normal speech production with five mid-sagittal slices.

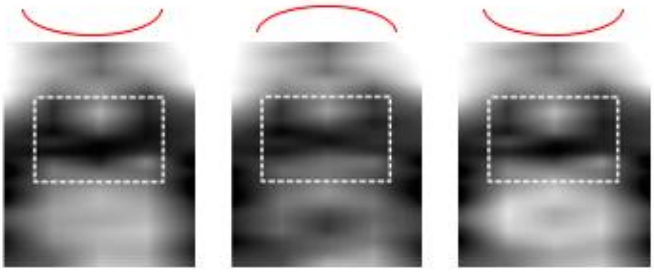
Figures 4.20 (a) – (c) further analyzes the contact between the tongue tip and the hard palate in both the mid-sagittal plane and the coronal plane. Figure 4.20 (a) depicts the location of the air stream in the vocal tract with a dotted line in the mid-sagittal plane. Figure 4.20 (b) depicts the formation of the air stream in the center of the tongue in /z/ of the /za/ sound. A different observation can be made from the /n/ of the /na/ sound when the tongue tip comes into complete contact with the hard palate. This can be further revealed by analyzing the coronal section posterior to the tongue tip, as can be seen in Fig. 4.20 (c). The curvature of the coronal tongue shape is describe with a red line on the top of the images. Obvious convexity in the tongue tip curvature is observed in /n/ of the /na/ sound while concavity is observed in /z/ of the /za/ sound.



(a)



(b)



(c)

Figure 4.20: (a) Location of the “tongue grooving” air stream in the vocal tract, (b) an air stream exists in the /za/ sound while disappears in the /na/ sound and (c) convexity and concavity of the coronal tongue shape.

CHAPTER 5

CONCLUSION

Dynamic speech imaging can suffer from low spatial and temporal resolution due to limited MR imaging speed. This thesis has proposed to use PS model-based approaches in the context of dynamic speech imaging.

This thesis has investigated the feasibility of applying the PS model to improve the spatiotemporal resolution of dynamic speech imaging. The PS model allows $(\boldsymbol{k}, t)$-space data to be sparsely sampled. The sparsely sampled data can be reconstructed with the PS model-based reconstruction methods to visualize small-size vocal articulators and fast-varying speech dynamics. Numerical simulations and *in vivo* experiments have demonstrated the effectiveness of the PS model to enable high spatiotemporal resolution dynamic speech imaging at high signal to noise ratio. Simulations and experiments also provide visualization of speech dynamics in multiple imaging planes to examine motion outside the midline of major articulators.

This thesis also investigates multiple choices of navigator sampling patterns. Specifically, the influence of the placement of the navigator sampling patterns on temporal dynamics of PS model-based reconstruction has been systematically studied. Although different navigator sampling patterns do not vary significantly in the several quantitative metrics considered, these patterns have been demonstrated to influence the detailed structure of articulator shaping and temporal features of speech dynamics. Specifically, non-Cartesian navigators yield better speech dynamics in terms of localized vocal tract shaping.

This thesis is the first attempt to apply a reconstruction method based joint PS and Sparse constraints to better capture speech dynamics. Specifically, PS-sparse has been employed to suppress the ill-conditioning issues associated with basic PS reconstruction and remove the spatiotemporal blurring associated with basic sparse reconstruction. The ability of PS-sparse in capturing better speech dynamics has been demonstrated in systematic simulations and preliminary *in vivo* experiments.

REFERENCES

[1] D. P. Kuehn, S. L. Ettema, M. S. Goldwasser, and J. C. Barkmeier, "Magnetic resonance imaging of the levator veli palatine muscle before and after primary palatoplasty," *Cleft Palate-Cran J.*, vol. 41, pp. 584-592, 2004.

[2] A. B. Lipira, L. M. Grames, D. Molter, D. Govier, and A. A. Kane, "Videofluoroscopic and nasendoscopic correlates of speech in velopharyngeal dysfunction," *Cleft Palate-Cran J.*, vol. 48, pp. 550-560, 2011.

[3] D. P. Kuehn, S. L. Ettema, M. S. Goldwasser, J. C. Barkmeier, and J. M. Wachtel, "Magnetic resonance imaging in the evaluation of occult submucous cleft palate*," Cleft Palate-Cran J.*, vol. 38, pp. 421-431, 2001.

[4] C. Filip, M. Matzen, I. Aagenæs, R. Aukner, L. Kjøll, H. E. Høgevold, F. Åbyholm, and K. Tønseth, "Speech and magnetic resonance imaging results following autologous fat transplantation to the velopharynx in patients with velopharyngeal insufficiency," *Cleft Palate-Cran J.*, vol. 4, pp. 708-716, 2011.

[5] S. R. Ventura, M. J. M. Vasconcelos, D. R. Freitas, I. M. Ramos, and J. M. R. S. Tavares, "Speech articulation assessment using dynamic magnetic resonance imaging techniques," in *Comp. Vis. Med. Img. Proc.*, 2011.

[6] J. Harrington and M. Tabain, *Towards a Better Understanding of Speech Production Processes*. New York: Psychology Press, 2004.

[7] A. G. Christodoulou, B. Zhao, H. Zhang, C. Ho, and Z.-P. Liang, "Four-dimensional MR cardiovascular imaging: Method and applications," in *Proc. IEEE Eng.Med. Bio. Conf.*, 2011, pp. 3732-3735.

[8] Z.-P. Liang, "Spatiotemporal imaging with partially separable functions," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2007, pp. 988-991.

[9] D. J. Pocci, “Analysis of Cartesian and projection ***(k**, t)*-space sampling patterns when using the partially separable functions model for cardiac perfusion MR imaging”, M. S. thesis, University of Illinois at Urbana-Champaign, May, 2010.

[10] C. Brinegar, H. Zhang, Y.-J. L.Wu, L. M. Foley, T. K. Hitchens, Q. Ye, C. Ho, and Z.-P. Liang, “First-pass perfusion cardiac MRI using the partially separable functions model with generalized support,” in *Proc. IEEE EMBC*, 2010, pp. 2833–2836.

[11] G. Bailly, P. Badin, D. Beautemps, and F. Elisei, "Speech technologies for augmented communication," http://hal.archives-ouvertes.fr/hal-00473026, Accessed Jan, 2012.

[12] B. E. Murdoch, J. V. Goozee, and L. M. L. Cahill, "Dynamic assessment of tongue function in children with dysarthria associated with acquired brain injury using electromagnetic articulography," in *2001 24th Annual Brain Impairment Conference*, vol. 2, pp. 63-65.

[13] J. O. Clarke, C. P. Gyawali, and R. P. Tatum, "High-resolution manometry," *Ann. NY Acad. Sci.*, vol. 1232, pp.349-357, 2011.

[14] M. Stone, "Laboratory techniques for investigating speech articulation, " *The Handbook of Phonetic Sciences*, Second Edition. Oxford, UK: Blackwell Publishing Ltd, 2010.

[15] F. G. Shellock, C. J. Schatz, P. M. Julien, J. M. Silverman, F. Steinberg, T. K. Foo, M. L. Hopp, and P. R. Westbrook, "Dynamic study of the upper airway with ultrafast spoiled GRASS MR imaging," *Am. J. Roentgenol.*, vol. 158, pp. 1019-1024, 1992.

[16] B. B. Wein, M. Drobnitzky, S. Klajman, W. Angerstein, "Evaluation of functional positions of tongue and soft palate with MR imaging: Initial clinical results," *J. Magn. Reson. Imaging*, vol. 1, pp. 381-383, 1991.

[17] A. Anagnostara, S. Stoeckli, O. M. Weber, and S. S. Kollias, "Evaluation of the anatomical and functional properties of deglutition with various kinetic high-speed MRI sequence," *J. Magn. Reson. Imaging*, vol. 14, pp. 194-199, 2001.

[18] A. J. Beer, P. Hellerhoff, A. Zimmermann, K. Mady, R. Sader, E. Rummeny, and C. Hannig, "Dynamic near-real-time magnetic resonance imaging for analyzing the velopharyngeal closure in comparison with videofluoroscopy," *J. Magn. Reson. Imaging*, vol. 20, pp. 791-797, 2004.

[19] S. Narayanan, A. Alwan, and K. Haker, "An articulatory study of fricative consonants using magnetic resonance imaging," *J Acoust Soc Am*, vol. 98, pp. 1325-1347, 1995.

[20] K. Kendall and R. Leonard, *Laryngeal Evaluation: Indirect Laryngoscopy to High-Speed*

Digital Imaging. New York: Thieme, 2010.

[21] S. Pilsworth, "Routine use of nasendoscopy to enhance the speech and language therapist's decision-making process in surgical voice restoration," *Otolaryngol Head Neck Surg*, vol. 145, pp. 86-90, 2011.

[22] Y. C. Chiang, F. P. Lee, C. L. Peng, and C. T. Lin, "Measurement of tongue movement during vowels production with computer-assisted B-mode and M-mode ultrasonography," *Otolaryngol Head Neck Surg*, vol. 128, pp. 805-814, 2003.

[23] A. J. Lundberg and M. Stone, "Three-dimensional tongue surface reconstruction: Practical considerations for ultrasound data," *J. Acoust. Soc. Am*, vol. 106, pp. 2858-2867, 1999.

[24] M. Stone, "A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data, " *J. Acoust. Soc. Am*, vol. 87, pp. 2207-2217, 1990.

[25] T. Bressmann, C.-L. Heng, and J. C. Irish, "Application of 2D and 3D ultrasound imaging in speech-language pathology, " *J. Speech Lang. Pathol. Audiol,* vol. 29, pp. 158-168.

[26] S. Kelly, K. M. Harris, E. Berry, J. Hutton, P. Roderick, J. Cullingworth, L. Gathercole and M. A. Smith, "A systematic review of the staging performance of endoscopic ultrasound in gastro-oesophageal carcinoma," *GUT*, vol. 49, pp. 534-539, 2001.

[27] W. L. Hardcastle and A. Marchal, *Speech Production and Speech Modelling.* Norwell: Kluwer Academic Publishers Group, 1990.

[28] K. Iskarous, " Patterns of tongue movement, " *Phonetica*, vol. 33, pp. 363-382, 2005.

[29] B. R. Pauloski, A. W. Rademaker, C. Lazarus, G. Boeckxstaens, P. J. Kahrilas, and J. A. Logemann, "Relationship between manometric and videofluoroscopic measures of swallow function in healthy adults and patients treated for head and neck cancer with various modalities," *Dysphagia*, vol. 24, pp. 196-203, 2009.

[30] T. R. Han, N. -J. Raik, J. W. Park, "Quantifying swallowing function after stroke: A functional dysphagia scale based on videofluoroscopic studies, " *Arch. Phys. Med. Rehabil*, vol. 82, pp. 677-682, 2005.

[31] K. Tom, I. R. Titze, E. A. Hoffman, and B. H. Story, "3-D vocal tract imaging and formant structure: Varying vocal register, pitch, and loudness," *J. Acoust. Soc. Am*, vol. 109, pp. 742-747.

[32] M. Yoshikawa, M. Yoshida, K. Tsuga, Y. Akagawa and M. Groher, "Comparison of three types of tongue pressure devices," *Dysphagia*, vol. 26, pp. 232-237, 2011.

[33] M. Sulter, D. G. Miller, R. F. Wolf, H. K. Schutte, H. P. Wit, and E. L. Mooyaart, "On the relation between the dimensions and resonance characteristics of the vocal tract: A study with MRI," *J. Magn. Reson. Imaging*, vol. 10, pp. 365-373, 1992.

[34] H. Sinagawa, T. Ono, E. Honda, S. Masaki, Y. Shimada, I. Fujimoto, T. Sasaki, A. Iriki and K. Ohyama, "Dynamic analysis of articulatory movement using magnetic resonance imaging movies: methods and implications in cleft lip and palate," *Cleft Palate Craniofacial J*, vol. 42, pp. 225-230, 2005.

[35] M. S. Inoue, T. Ono, E. Honda, and T. Kurabayashi, "Application of magnetic resonance imaging movie to assess articulatory movement," *Orthod Craniofacial Res*, vol. 9, pp. 157-162, 2006.

[36] M. S. NessAiver, M. Stone, V. Parthasarathy, Y. Kahana, A. Kots and A. Paritsky, "Recording high-quality speech during tagged cine-MRI studies using a fiber optic microphone," *J Magn Reson Imag*, vol. 23, pp. 92-97, 2006.

[37] A. Haase, J. Frahm, D. Matthaei, W. H ̈anicke, and K.-D. Merboldt, “FLASH imaging: rapid NMR imaging using low flip-angle pulses,” *J. Magn. Reson.*, vol. 67, pp. 258-266,1986.

[38] J. A. Kim, V. R. Narra, “Magnetic resonance imaging with true fast imaging with steady-state precession and half-Fourier acquisition single-shot turbo spin-echo sequences in cases of suspected placenta accrete,” *Acta Radiol.*, vol. 45,pp. 692-698, 2004.

[39] R. Turner, D. L. Bihan, J. Maier, R. Vavrek, L. K. Hedges, and J. Pekar, “Echo-Planar Imaging of intravoxel incoherent motion,” *Radiol.*, vol.177,pp.407-414, 1990.

[40] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, “SENSE: Sensitivity encoding for fast MRI,” *Magn. Reson. Med.*, vol. 42, pp. 952-962, 1999.

[41] D. K. Sodickson and W. J. Manning, “Simultaneous acquisition of spatial harmonics (SMASH): Fast imaging with radiofrequency coil arrays,” *Magn. Reson. Med.*, vol. 38, pp. 591-603, 1997.

[42] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase, “Generalized autocalibrating partially parallel acquisitions (GRAPPA),” *Magn. Reson. Med.*, vol. 47, pp. 1202-1210, 2002.

[43] B. P. Sutton, C. A. Conway, Y. Bae, R. Seethamraju, and D. P. Kuehn, "Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (FLASH) at 3T," *J. Magn. Reson. Imaging*, vol. 32, pp. 1228-1237, 2010.

[44] B. P. Sutton, C. Conway, Y. Bae, C. Brinegar, Z.-P. Liang, and D. P. Kuehn, "Dynamic imaging of speech and swallowing with MRI," 31st Ann. Conf. IEEE EMBS, September 2009, pp. 6651-6654.

[45] K. Mady, and A. Beer, "A real-time MRI evaluation of consonant production after oral tumour surgery," *Grazer Linguistische Studien,* vol. 62, pp. 77-94.

[46] Y.-C. Kim, M. L. Proctor, S. S. Narayanan, and K. S. Nayak, "Refined imaging of lingual articulation using real-time multislice MRI," *J. Magn. Reson. Imaging*, online version.

[47] Q.-S. Xiang and R. M. Henkelman, "K-space description for MR imaging of dynamic objects," *Magn. Reson. Med.*, vol. 29, pp. 422-428, 1993.

[48] M. A. Bernstein, K. F. King, and X. J. Zhou, *Handbook of MRI Pulse Sequences*. Burlington: Elsevier Academic Press, 2004.

[49] M. Blaimer, F. Breuer, M. Mueller, R. M. Heidemann, M. A. Griswold, and P. K. Jakob, "SMASH, SENSE, PILS, GRAPPA How to choose the optimal method," *Top. Magn. Reson. Imaging*,vol.15, pp. 223-236, 2004.

[50] B. Madore, G. H. Glover, and N. J. Pelc, "Un-aliasing by Fourier-encoding the overlaps using the temporal dimension (UNFOLD), applied to cardiac imaging and fMRI," *Magn. Reson. Med.*, vol. 42, pp. 813-828, 1999.

[51] P. Kellman, F. H. Epstein, and E. R. McVeigh, "Adaptive sensitivity encoding incorporating temporal filtering (TSENSE)," *Magn. Reson. Med.*, vol. 45, pp. 846-852, 2001.

[52] N. Aggarwal and Y. Bresler, "Patient-adapted reconstruction and acquisition dynamic imaging method (PARADIGM) for MRI," *Inverse Probl.*, vol. 24, p. 045015, 2008.

[53] Y. Bresler and S. P. Litke, "A parametric technique for superresolution image reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1990, pp. 1205-1208.

[54] J. Tsao, "k-t SENSE for efficient dynamic parallel imaging by joint space-time consideration," www.mr.ethz.ch/parallelmri04/abstracts/pub/Tsao.pdf, Accessed May, 2010.

[55] J. Tsao, P. Boesiger, and K. P. Pruessmann, "*k-t* BLAST and *k-t* SENSE: Dynamic MRI with high frame rate exploiting spatiotemporal correlations," *Magn. Reson. Med.*, vol. 50, pp. 1031-1042, 2003.

[56] H. Jung, K. Sung, K. S. Nayak, E. Y. Kim, and J. C. Ye, "k-t FOCUSS: A general compressed sensing framework for high resolution dynamic MRI," *Magn. Reson. Med.*, vol. 61, pp. 103-116, 2009.

[57] M. Lustig, J. M. Santos, D. L. Donoho, and J. M. Pauly, "k-t SPARSE: High frame rate dynamic MRI exploiting spatio-temporal sparsity," in *Proc. Int. Soc. Magn. Reson. Med.*, 2006, p. 2420.

[58] Z.-P. Liang, F. Boada, T. Constable, E. M. Haacke, P. C. Lauterbur, and M. R. Smith, "Constrained reconstruction methods in MR imaging," *Rev. Magn. Reson. Med.*, vol. 4, pp. 67-185, 1992.

[59] B. Zhao, J. P. Haldar, and Z.-P. Liang, "PSF model-based reconstruction with sparsity constraint: Algorithm and application to real-time cardiac MRI," in *Proc. IEEE Eng. Med. Bio. Conf.*, 2010, pp. pp. 3390-3393.

[60] B. Madore, "Using UNFOLD to remove artifacts in parallel imaging and in partial-fourier Imaging," *Magn. Reson. Med.*, vol. 48, pp. 493-501, 2002.

[61] J. Tsao, "On the UNFOLD method," *Magn. Reson. Med.*, vol. 47, pp. 202-207, 2002.

[62] B. Sharif, "Distortion-optimal parallel MRI with sparse sampling: from adaptive spatiotemporal acquisition to self-calibrating reconstruction," Ph.D. dissertation, University of Illinois at Urbana-Champaign, Aug. 2010.

[63] J. Tsao, K. Pruessmann, and P. Boesiger, "Prior-information-enhanced dynamic imaging using single or multiple coils with k-t BLAST and k-t SENSE," in *Proc. Int. Soc. Magn. Reson. Med.*,2002, pp. 10.

[64] H. Jung, J. Park, J. Yoo, and J. C. Ye, "Radial k-t FOCUSS for high-resolution cardiac cine MRI," *Magn. Reson. Med.*, vol. 63, pp. 68-78, 2010.

[65] D. Donoho, and Y. Tsaig, "Fast resolution of l1-norm minimization problems when the solution may be sparse," *Bernoulli*, vol. 54, pp. 1-44, 2006.

[66] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems," *Proc. IEEE*, vol. 98, pp. 948-958, 2010.

[67] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, pp. 707-710, 2007.

[68] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, pp. 471-501, 2010.

[69] R. A. DeVore, "Deterministic constructions of compressed sensing matrices," *J. Complexity*, vol. 23, pp. 918-925, 2007.

[70] R.M.Willett and M. Raginsky, "Performance bounds on compressed sensing with Poisson noise," in *Int. Symp. Inf. Theory*, vol. 2, pp. 174 - 178, 2009.

[71] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed Sensing MRI," *IEEE Signal Process. Mag.*, vol. 27, pp. 72-82, 2008.

[72] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted norm minimization algorithm," *IEEE T. Signal Process.*, vol. 45, no. 3, pp. 600-616, 1997.

[73] H. Jung, J. Yoo, and J. C. Ye, "Generalized k-t BLAST and k-t SENSE using FOCUSS," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2007, pp. 145-148.

[74] M. Wang, W. Xu, and A. Tang, "On the performance of sparse recovery via *lp*-minimization ($0 \le p \le 1$)," *IEEE T. Inf. Theory*, vol. 57, no. 11, pp. 7255-7278, 2011.

[75] J. P. Haldar, and Z.-P. Liang, "Low-rank approximations for dynamic imaging," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2011,pp. 052-1055.

[76] B. Zhao, J. P. Haldar, C. Brinegar, and Z.-P. Liang, "Low rank matrix recovery for real-time cardiac MRI," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2010, pp. 996-999.

[77] A. G. Christodoulou, S. D. Babacan, and Z.-P. Liang, "Accelerating cardiovascular imaging by exploiting regional low-rank structure via group sparsity," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2012.

[78] A. G. Christodoulou, B. Zhao, Z.-P. Liang, "Regularized image reconstruction for PS model-based cardiovascular MRI," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2011, pp. 57-60.

[79] H. Nguyen, X. Peng, M. Do, and Z-P. Liang, "Spatiotemporal denoising of MR spectroscopic imaging data by low-rank approximation," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2011, pp. 857-860.

[80] H. Nguyen, J. P. Haldar, M. Do, and Z-P. Liang, "Denoising of MR spectroscopic imaging data with spatial-spectral regularization," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2011, pp. 720-723.

[81] J. P. Haldar, "Constrained imaging: denoising and sparse sampling," Ph.D. dissertation, University of Illinois at Urbana-Champaign, May. 2011.

[82] Z.-P. Liang, H. Jiang, C. P. Hess, and P. C. Lauterbur, "Dynamic imaging by model estimation," *Int. J. Imag. Syst. Tech.*, vol. 8, pp. 551-557, 1997.

[83] G. C. Tiao, and G.E.P. Box, "Modeling multiple time series with applications," *J. Amer. Statistical Assoc.*, vol.76, pp. 802-816, 1981.

[84] C. Brinegar, H. Zhang, Y.-J.L. Wu, L. M. Foley, T. K. Hitchens, Q. Ye, C. Ho, and Z.-P. Liang, "Real-time cardiac MRI using prior spatial-spectral information," in *Proc. IEEE Eng. Med. Bio. Conf.*, 2009, pp. 4383-4386.

[85] B. Zhao, J. P. Haldar, and Z.-P. Liang, "PSF model-based reconstruction with sparsity constraint: Algorithm and application to real-time cardiac MRI," in *Proc. IEEE Eng. Med. Bio. Conf.*, 2010, pp. pp. 3390-3393.

[86] B. Sutton, C. Conway, Y. Bae, C. Brinegar, Z.-P. Liang, and D. P. Kuehn, "Dynamic imaging of speech and swallowing with MRI," in *Proc. IEEE Eng. Med. Bio. Conf.*, 2009, pp. 6651-6654.

[87] A. G. Christodoulou, C. Brinegar, J. P. Haldar, H. Zhang, Y.-J. L. Wu, L. M. Foley, T. K. Hitchens, Q. Ye, C. Ho, and Z.-P. Liang, "High-resolution cardiac MRI using partially separable functions and weighted spatial smoothness regularization," in *Proc. IEEE Eng. Med. Bio. Conf.*, 2010, pp. 883-886.

[88] X. Qu, X. Cao, D. Guo, C. Hu, Z. Chen, "Compressed sensing MRI with combined sparsifying transforms and smoothed l_0 norm minimization," in *Proc. Intl. Conf. Acoust. Speech Signal Process.*, 2010.

[89] M. Fu, A. G. Christodoulou, A. T. Naber, D. P. Kuehn, Z.-P. Liang, and B. P. Sutton, "High-frame-rate multislice speech imaging with sparse sampling of ***(k**, t)*-space," in *Proc. 12th Ann. Intl. Soc. Magn. Res. Imag.*, 2012.

[90] A. Majumdar, R. K. Ward, "An algorithm for sparse MRI reconstruction by Schatten *p*-norm minimization," *Magn. Reson. Imag.*, vol. 29, pp. 408-417, 2011.

[91] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE T. Inform. Theory*, 2005, pp. 4203-4215.

[92] B. Zhao, J. P. Haldar, A. G. Christodoulou, and Z.-P. Liang, "Image reconstruction from highly undersampled ***(k**, t)*-space data with joint partial separability and sparsity constraints," submitted to *IEEE T. Med. Imag.*, 2012.

[93] J. Lim, and M.-H. Yang, "A direct method for modeling non-rigid motion with thin plate spline," in *Proc. CVPR*, 2005, vol. 1, pp. 1196-1202.

[94] G. Donato, and S. Belongie, "Approximate Thin Plate Spline Mappings," in *Proc. Seventh*

European Conf. Computer Vision, vol. 3, pp. 21-31, 2002.

[95] D. C. Adams, "Methods for shape analysis of control data from articulated structures," *Evolutionary Ecology Research*, vol. 1, pp. 959-970, 1999.

[96] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations, " *IEEE T. Pattern Anal. Machine Intell.*, vol. 11, pp.567 - 585 , 1989.

[97] S. Wang, J. X. Ji and Z.-P. Liang, "Control-based shape deformation with topology-preserving constraints, " in *Proc. 9th IEEE Int. Conf. Computer Vision*, vol. 2, pp.923, 2003.

[98] J. A. d'Arcy, D. J. Collins, I. J. Rowland, A. R. Padhani, and M. O. Leach, "Applications of sliding window reconstruction with Cartesian sampling for dynamic contrast enhanced MRI, " *NMR Biomed.*, vol. 15, no. 2, pp.174 - 183 , 2002